

LREC-COLING 2024

**11th Workshop on the  
Representation and Processing of Sign Languages:  
Evaluation of Sign Language Resources  
([sign-lang@LREC-COLING 2024](mailto:sign-lang@LREC-COLING 2024))**

Workshop Proceedings

Editors

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke,  
Julie A. Hochgesang, Johanna Mesch, and Marc Schulder

25 May 2024  
Torino, Italia

**Proceedings of the LREC-COLING 2024 11th Workshop on the  
Representation and Processing of Sign Languages:  
Evaluation of Sign Language Resources  
(sign-lang@LREC-COLING 2024)**

Copyright ELRA Language Resources Association (ELRA), 2024  
These proceedings are licensed under a Creative Commons  
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-30-2  
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association  
and the International Committee on Computational Linguistics



## Preface

This collection of papers stems from the 11th Workshop on the Representation and Processing of Sign Languages which takes place as a satellite workshop to the LREC-COLING 2024 Joint Conference in Turin, Italy.

While there has been occasional attention to sign languages at the main LREC conference, the focus there is on spoken languages in their written and spoken forms. This series of workshops, however, offers a forum for researchers focussing on sign languages, especially on corpus data and corpus technology for sign languages.

This year's hot topic "Evaluation of Sign Language Resources" addresses the challenge that as the field is maturing, it becomes increasingly important to assess the quality of sign language resources for a large variety of tasks. This relates to both automatic and human-based evaluation procedures and to a large variety of sign language resources and tools.

The contributions composing this volume are presented in alphabetical order by the first author. For the reader's convenience, an author index is provided as well.

Once again, we would like to thank all members of the program committee who helped us tremendously by reviewing the submissions to the workshop within a very short timeframe!

Finally, we would like to point the reader to the sign-lang@LREC Anthology at

<https://www.sign-lang.uni-hamburg.de/lrec/>

The anthology contains all publications of the workshop series as well as sign language papers from the LREC main conference and its other workshops. It offers author and topic indices across all papers, stable URLs for all workshop papers and their supplementary materials, as well as bibliographical (BibTeX) data for all entries. Happy browsing!

The Editors

# Organizers

## Organizing Committee

Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece  
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece  
Thomas Hanke, University of Hamburg, Hamburg, Germany  
Julie A. Hochgesang, Gallaudet University, Washington, USA  
Johanna Mesch, Stockholm University, Stockholm, Sweden  
Marc Schulder, University of Hamburg, Hamburg, Germany

## Program Committee

Sam Bigeard, Inria Centre, Nancy, France  
Penny Boyes-Braem, University of Zurich, Zurich, Switzerland  
Richard Bowden, University of Surrey, Guildford, United Kingdom  
Annelies Braffort, CNRS/LISN, Orsay, France  
Carl Börstell, University of Bergen, Bergen, Norway  
Kearsy Cormier, University College London, London, United Kingdom  
Athanasia-Lida Dimou, Institute for Language and Speech Processing, Athens, Greece  
Sarah Ebling, University of Zurich, Zurich, Switzerland  
Eleni Efthimiou, Institute for Language and Speech Processing, Athens, Greece  
Michael Filhol, CNRS/LISN, Orsay, France  
Stavroula-Evita Fotinea, Institute for Language and Speech Processing, Athens, Greece  
Kathleen Currie Hall, University of British Columbia, Vancouver, Canada  
Thomas Hanke, University of Hamburg, Hamburg, Germany  
Julie A. Hochgesang, Gallaudet University, Washington, USA  
Amy Isard, University of Hamburg, Hamburg, Germany  
Vadim Kimmelman, University of Bergen, Bergen, Norway  
Reiner Konrad, University of Hamburg, Hamburg, Germany  
Maria Kopf, University of Hamburg, Hamburg, Germany  
Anna Kuder, University of Cologne, Cologne, Germany  
Gabriele Langer, University of Hamburg, Hamburg, Germany  
Özge Mercanoğlu Sincan, University of Surrey, Guildford, United Kingdom  
Johanna Mesch, Stockholm University, Stockholm, Sweden  
Hope E. Morgan, University of Hamburg, Hamburg, Germany  
Carol Neidle, Boston University, Boston, USA  
Justin Power, University of Texas, Austin, USA  
Paweł Rutkowski, University of Warsaw, Warsaw, Poland  
Marc Schulder, University of Hamburg, Hamburg, Germany  
Tsubasa Uchida, NHK Science & Technology Research Laboratories, Tokyo, Japan  
Sabrina Wähl, University of Hamburg, Hamburg, Germany  
Joanna Wójcicka, University of Warsaw, Warsaw, Poland  
Rosalee Wolfe, Institute for Language and Speech Processing, Athens, Greece

## Table of Contents

<i>Advancing Annotation for Continuous Data in Swiss German Sign Language</i> Alessia Battisti, Katja Tissi, Sandra Sidler-Miserez and Sarah Ebling .....	1
<i>Person Identification from Pose Estimates in Sign Language</i> Alessia Battisti, Emma van den Bold, Anne Göhring, Franz Holzknicht and Sarah Ebling .....	13
<i>Data Integration, Annotation, and Transcription Methods for Sign Language Dialogue with Latency in Videoconferencing</i> Mayumi Bono, Tomohiro Okada, Victor Skobov and Robert Adam .....	26
<i>Evaluating the Alignment of Utterances in the Swedish Sign Language Corpus</i> Carl Börstell .....	36
<i>How to Approach Lexical Variation in Sign Language Corpora</i> Carl Börstell .....	46
<i>Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas</i> Aashaka Desai, Maartje De Meulder, Julie A. Hochgesang, Annemarie Kocab and Alex X. Lu .....	54
<i>Evaluating Inter-Annotator Agreement for Non-Manual Markers in Sign Languages</i> Lyke D. Esselink, Marloes Oomen and Floris Roelofsen .....	66
<i>A software editor for the AZVD graphical Sign Language representation system</i> Michael Filhol and Thomas von Ascheberg .....	77
<i>Content Questions in Sign Language – From theory to language description via corpus, experiments, and fieldwork</i> Robert Gavrilesco, Carlo Geraci and Johanna Mesch .....	86
<i>Matignon-LSF: a Large Corpus of Interpreted French Sign Language</i> Julie Halbout, Diandra Fabre, Yanis Ouakrim, Julie Lascar, Annelies Braffort, Michèle Gouiffès and Denis Beautemps .....	95
<i>Phonological Transcription of the Canadian Dictionary of ASL as a Language Resource</i> Kathleen Currie Hall, Anushka Asthana, Maggie Reid, Yiran Gao, Grace Hobby, Oksana Tkachman and Kaili Vesik .....	102
<i>Retrospective of Kazakh-Russian Sign Language Corpus Formation</i> Alfarabi Imashev .....	111
<i>Enhancing Syllabic Component Classification in Japanese Sign Language by Pre-training on Non-Japanese Sign Language Data</i> Jundai Inoue, Makoto Miwa, Yutaka Sasaki and Daisuke Hara .....	124
<i>Building Your Query Step by Step: A Query Wizard for the MY DGS – ANNIS Portal of the DGS Corpus</i> Amy Isard .....	132

<i>Investigating Motion History Images and Convolutional Neural Networks for Isolated Irish Sign Language Fingerspelling Recognition</i> Hafiz Muhammad Sarmad Khan, Irene Murtagh and Simon D. McLoughlin .....	141
<i>Shedding Light on the Underexplored: Tackling the Minor Sign Language Research Topics</i> Jung-Ho Kim, Changyong Ko, Mathew Huerta-Enochian and Seung Yong Ko .....	148
<i>Headshakes in NGT: Relation between Phonetic Properties &amp; Linguistic Functions</i> Vadim Kimmelman, Marloes Oomen and Roland Pfau .....	160
<i>Nonmanual Marking of Questions in Balinese Homesign Interactions: a Computer-Vision Assisted Analysis</i> Vadim Kimmelman, Ari Price, Josefina Safar, Connie de Vos and Jan Bulla .....	169
<i>An Extension of the NGT Dataset in Global Signbank</i> Ulrika Klomp, Lisa Gierman, Pieter Manders, Ellen Nauta, Gomèr Otterspeer, Ray Pelupessy, Galya Stern, Dalene Venter, Casper Wubbolts, Marloes Oomen and Floris Roelofsen .....	179
<i>Corpus à la carte – Improving Access to the Public DGS Corpus</i> Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken and Oliver Böse .....	185
<i>Introducing the DW-DGS – The Digital Dictionary of DGS</i> Gabriele Langer, Anke Müller, Sabrina Wähl, Felicitas Otte, Lea Sepke and Thomas Hanke .....	195
<i>Annotation of LSF subtitled videos without a pre-existing dictionary</i> Julie Lascar, Michèle Gouiffès, Annelies Braffort and Claire Danet .....	205
<i>Capturing Motion: Using Radar to Build Better Sign Language Corpora</i> Evie Malaia, Joshua Borneman and Sevgi Gurbuz .....	214
<i>Exploring Latent Sign Language Representations with Isolated Signs, Sentences and In-the-Wild Data</i> Fredrik Malmberg, Anna Klezovich, Johanna Mesch and Jonas Beskow .....	220
<i>Quantitative Analysis of Hand Locations in both Sign Language and Non-linguistic Gesture Videos</i> Niels Martínez-Guevara and Arturo Curiel .....	226
<i>Formal Representation of Interrogation in French Sign Language</i> Emmanuella Martinod and Michael Filhol .....	236
<i>Multilingual Synthesis of Depictions through Structured Descriptions of Sign: An Initial Case Study</i> John McDonald, Eleni Efthimiou, Stavroula-Evita Fotinea and Rosalee Wolfe .....	245
<i>Swedish Sign Language Resources from a User’s Perspective</i> Johanna Mesch, Thomas Björkstrand, Eira Balkstam, Patrick Hansson and Nikolaus Riemer Kankkonen .....	255
<i>Sign Language Translation with Gloss Pair Encoding</i> Taro Miyazaki, Sihan Tan, Tsubasa Uchida and Hiroyuki Kaneko .....	263

<i>SignCollect: A ‘Touchless’ Pipeline for Constructing Large-scale Sign Language Repositories</i>	
Gomèr Otterspeer, Ulrika Klomp and Floris Roelofsen .....	270
<i>The EASIER Mobile Application and Avatar End-User Evaluation Methodology</i>	
Frankie Picron, Davy Van Landuyt, Rehana Omardeen, Eleni Efthimiou, Rosalee Wolfe, Stavroula-Evita Fotinea, Theodore Goulas, Christian Tismer, Maria Kopf and Thomas Hanke .....	277
<i>VisuoLab: Building a sign language multilingual, multimodal and multifunctional platform</i>	
Christian Rathmann, Ronice Muller de Quadros, Thomas Geißler, Christian Peters, Francisco Fernandes, Milene Peixer Loio and Diego França .....	283
<i>3D-LEX v1.0 – 3D Lexicons for American Sign Language and Sign Language of the Netherlands</i>	
Oline Ranum, Gomèr Otterspeer, Jari I. Andersen, Robert G. Belleman and Floris Roelofsen .....	291
<i>Signbank 2.0 of Sign Languages: Easy to Administer, Easy to Use, Easy to Share</i>	
Ronice Muller de Quadros, Christian Rathmann, Peter Zalán Romanek, Francisco Fernandes and Sther Condé .....	303
<i>STK LSF: A Motion Capture Dataset in LSF for SignToKids</i>	
Clément Reverdy, Sylvie Gibet and Thibaut Le Naour .....	316
<i>Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition</i>	
Kyunggeun Roh, Huije Lee, Eui Jun Hwang, Sukmin Cho and Jong C. Park .....	324
<i>Decoding Sign Languages: The SL-FE Framework for Phonological Analysis and Automated Annotation</i>	
Karahan Şahin and Kadir Gökgöz .....	336
<i>Signs and Synonymity: Continuing Development of the Multilingual Sign Language Wordnet</i>	
Marc Schulder, Sam Bigeard, Maria Kopf, Thomas Hanke, Anna Kuder, Joanna Wójcicka, Johanna Mesch, Thomas Björkstrand, Anna Vacalopoulou, Kyriaki Vasilaki, Theodore Goulas, Stavroula-Evita Fotinea and Eleni Efthimiou .....	344
<i>Facial Expressions for Sign Language Synthesis using FACSHuman and AZee</i>	
Paritosh Sharma, Camille Challant and Michael Filhol .....	355
<i>Eye Blink Detection in Sign Language Data Using CNNs and Rule-Based Methods</i>	
Margaux Susman and Vadim Kimmelman .....	362
<i>SEDA: Simple and Effective Data Augmentation for Sign Language Understanding</i>	
Sihan Tan, Taro Miyazaki, Katsutoshi Itoyama and Kazuhiro Nakadai .....	371
<i>HamNoSys-based Motion Editing Method for Sign Language</i>	
Tsubasa Uchida, Taro Miyazaki and Hiroyuki Kaneko .....	377

<i>SignaMed: a Cooperative Bilingual LSE-Spanish Dictionary in the Healthcare Domain</i> Manuel Vázquez-Enríquez, José Luis Alba-Castro, Ania Pérez-Pérez, Carmen Cabeza-Pereiro and Laura Docío-Fernández .....	387
<i>Diffusion Models for Sign Language Video Anonymization</i> Zhaoyang Xia, Yang Zhou, Ligong Han, Carol Neidle and Dimitris N. Metaxas .....	396
<i>A Multimodal Spatio-Temporal GCN Model with Enhancements for Isolated Sign Recognition</i> Yang Zhou, Zhaoyang Xia, Yuxiao Chen, Carol Neidle and Dimitris N. Metaxas .....	409

# Advancing Annotation for Continuous Data in Swiss German Sign Language

Alessia Battisti\* , Katja Tissi† , Sandra Sidler-Miserez†, Sarah Ebling\* 

\*University of Zurich  
Andreasstrasse 15, 8050 Zurich  
{battis, ebling}@cl.uzh.ch

†University of Teacher Education in Special Needs  
Schaffhauserstrasse 239, 8050 Zurich  
katja.tissi@hfh.ch | sandysidler@gmail.com

## Abstract

This paper presents a transcription and annotation scheme introduced specifically for L1 and L2 continuous data of Swiss German Sign Language, with potential applicability to other sign languages. The scheme includes a novel way of annotating linguistic errors in L2 data, thereby contributing to a deeper understanding of sign language learning. An initial validation approach is outlined, revealing challenges and underscoring the necessity for a more comprehensive method for validating sign language (learner) data. The paper emphasizes the overarching goal of achieving interoperability among sign language corpora and research groups, particularly in advancing sign language data validation techniques.

**Keywords:** Sign language data, learner corpus, annotation scheme, inter-annotator agreement

## 1. Introduction

Transcribing and annotating sign language data represents a significant bottleneck in the development of sign language corpora, especially when aiming for substantially sized, well-annotated datasets for automated Sign Language Processing (SLP) tasks. Many challenges in SLP arise not only due to a scarcity of consistent and detailed annotations but also due to the variation in annotation standards and granularity across projects.

In sign language corpus creation, it is crucial for annotation schemes and guidelines to adopt a broader perspective, characterized as “holistic and forward-thinking” by Hodge and Crasborn (2022). In a “holistic” approach, both basic and detailed annotations are combined from the beginning of the annotation process. The former, comparable to transcription (Konrad, 2011), includes segmentation and tokenization, which involves identifying manual actions, usually at the level of lexical units. The latter enriches the transcription with a more detailed level of annotation, such as non-manual actions and potentially grammatical functions. A more comprehensive approach such as this promotes best practices and represents a step towards standardization of signed language corpora.

This paper presents the development of an annotation scheme integrating basic and detailed annotations, designed for multidisciplinary use in sign language linguistics, automatic sign language assessment, and SLP. The development of this scheme was an integral part in constructing a longitudinal corpus of Swiss German Sign Lan-

guage (*Deutschschweizerische Gebärdensprache*, DSGS) second language (L2) learners, alongside a corpus of native/early learners (L1) of DSGS.

We summarize the process of annotating sign language (learner) data and present the annotation scheme. Given that the data is continuous signing that exceeds the level of individual signs, our scheme primarily focuses on the annotation of non-manual components that sometimes stretch across multiple manual signs. Furthermore, we address the annotation of L2 errors and suggest the potential of our scheme for future annotation of sign language (learner) data to enhance interoperability of datasets and thus facilitate cross-linguistic studies. Finally, we introduce an initial validation approach and preliminary results, highlighting the challenges encountered and the need for a comprehensive validation method for sign language (learner) data.

Section 2 introduces previous work in the area of sign language annotation, with a focus on inter-annotator agreement in sign language data. Section 3 summarizes our annotator process, while Section 4 describes the annotation scheme in details. In Section 5, we outline an initial validation approach on our annotated data.

## 2. Related Work

### 2.1. Annotations of Sign Language (Learner) Data

Several attempts have been made to define standards and best practices in sign language data



annotation (Nonhebel et al., 2004; Johnston, 2010; Schembri and Crasborn, 2010; Cormier et al., 2016). The selection of annotation scheme and the specificity of its labels are frequently influenced by the linguistic theories embraced by the researchers and by their research questions (Hodge and Crasborn, 2022). For instance, lexical frequency and morphosyntactic analysis guide the annotation scheme for the Auslan Corpus (*Australian Sign Language*) (Johnston, 2008), while phonetics and phonology shape the scheme for the NGT Corpus (*Nederlandse Gebarentaal*, Sign Language of the Netherlands; Crasborn et al., 2006-2017).

Kopf et al. (2022) delineates commonalities and differences between annotation conventions as applied to several publicly accessible sign language corpora. In the section dedicated to non-manual components, the authors point out that there are few studies describing the annotation of non-manual activities. Among the most recent works, Johnston (2019) provides detailed insights into the considerations made to annotate the form and the function of these components in Auslan, while Wallin and Mesch (2018) describe how they treated and annotated these activities in the corpus of Swedish Sign Language (*Svenskt teckenspråk*, STS).

Given the importance of these components at the sentence and discourse levels, Gabarró-López and Meurant (2014) explain how to use certain non-manual components, including head nod or movement, eye blink, and gaze, as criteria to facilitate sign language discourse segmentation in French Belgian Sign Language (*Langue des signes de Belgique francophone*, LSF). Similarly, to describe the components' function at the sentence and discourse levels, Lackner (2019) illustrates their annotation and their potential configurations in Austrian Sign Language (*Österreichische Gebärdensprache*, ÖGS).

However, none of the aforementioned studies specifically address the annotation of manual and non-manual components in sign language learner data. Despite the increased interest in research focusing on sign second language acquisition (SSLA) and the creation of datasets from non-native signers (L2 signers) (Schönström, 2021), management and annotation of L2 data remains an understudied area (Mesch and Schönström, 2018). This is characterized by a lack of guidelines for annotating errors or L2 linguistic structures. In addition to basic or detailed annotations similar to those applied to L1 data, L2 data is typically enriched with annotations that highlight deviations from canonical forms or disfluencies, a common practice also employed in the studies of spoken language learning (Gilquin and De Cock, 2011).

For analyzing the Corpus in Swedish Sign Lan-

guage as a Second Language (SSL-L2), Mesch and Schönström (2018) proposed a method to annotate typical L2 structures, which includes conventions for annotating phenomena specific to L2 languages. The authors build upon their previous studies on annotations of non-manual components and errors (Schönström and Mesch, 2014; Mesch et al., 2016).

Until recently, research on SSLA has primarily focused on analyzing individual glosses and manual errors (Rosen, 2004; Ortega and Morgan, 2015; Ebling et al., 2021; Kurz et al., 2023). However, there has been a growing interest in investigating higher-level linguistic constructions, such as sentences or discourse, highlighting the need for annotating non-manual components also for L2. For example, Mesch and Schönström (2020) explored the use of mouth actions in SSL-L2, while Gulamani et al. (2020) examined the adoption of different viewpoints in British Sign Language (BSL) learners.

## 2.2. Inter-annotator Agreement in Sign Language Data

None of the above-mentioned studies present an approach for the validation of annotated data. Studies on sign languages either do not report on reliability or provide only superficial ratings of inter-rater agreement (Schembri and Crasborn, 2010). For example, Hodge (2014) conducted a thorough examination of the annotation procedure, where additional annotators reviewed annotations of clause-like expressions by way of re-analysis.

Calculating agreement on sign language data annotations is a complex process that must consider multiple variables, such as the diversity of time spans and labels used.

In the context of annotations on behavioral studies, Andersson and Sandgren (2016) proposed a method called *temporally weighted overlap ratio*, to use with the ELAN annotation software (Wittenburg et al., 2006), to calculate agreement between two annotated events. Considering a certain time span, the authors search for an event in two different annotation transcripts. If an event is found and has the same label for Annotator A and Annotator B, an agreement is calculated based on the time overlap between the two events weighted by the maximum length of the event. This approach can also be applied to measure agreement between two events in a given time span in sign language data.

## 3. Annotation Process

As mentioned in Section 1, we devised the annotation process and scheme as part of constructing a longitudinal corpus of continuous DSGS L2 pro-



duction, in parallel with an L1 control corpus. In total, 35 participants were recorded, resulting in approximately 70 hours of recorded data.

The L1 control corpus comprises recordings of ten deaf signers performing the same tasks as the DSGS learners. Examples of tasks include picture or video retelling. We enlisted deaf signers who use DSGS as their primary language and acquired the language at different ages ( $M=3.8$ ,  $SD=6.1$ ). Among the 25 L2 participants, 14 were students of a DSGS interpreter training program. We followed these students throughout their language learning journey by recording their language production four times over an 18-month period.

Annotation is carried out by a team comprising two L1 deaf expert annotators with extensive experience in teaching and researching sign language, alongside two L1 deaf annotators-in-training, all of whom are project members. The data is annotated using the iLex software (Hanke and Storz, 2008), allowing for the linkage of all sign tokens in the corpus to their corresponding sign types in the lexicon and propagating any changes to sign types across all transcripts.

Figure 1 illustrates the data processing steps, starting from raw data in the recording phase to the subsequent data annotation rounds. Initially, we pre-process the data and generate transcripts that include selected tiers for both manual and non-manual components, with task boundaries automatically annotated based on recording software timestamps.

The data then undergoes two main rounds of processing. The first round involves segmenting tasks into sentences and sign units, identifying manual and non-manual components for both L1 and L2 data, and labeling the time span for each identified feature. In the second round, deviations from the canonical form are identified and labeled in the L2 data. Additional tiers are added to the L2 transcripts to facilitate marking deviations for both manual and non-manual components. A third round involves cross-checking and validating annotated data applying the four-eyes principle, where 20% of annotated data are re-annotated by the two expert annotators to calculate agreement. Annotations by annotators-in-training undergo double-checking, with corrections made as needed. Disagreements between annotators are discussed with an expert sign language linguist to understand the disagreement factors and resolve differences.

Due to the comprehensive nature of the annotation task and the corpus’s extensive volume, only selected tasks of the first two data collection points have been annotated thus far. On average, for both L1 and L2 data, annotators require 30 minutes to annotate a sentence containing six glosses.

Figure 2 displays a sample transcript in iLex for

an L2 learner production, showing annotations from the first and second rounds.

## 4. Annotation Scheme

In developing the annotation scheme, we were faced with the challenge of determining the granularity of the annotation, which is dependent upon the intended application of the corpus.

In our scheme, we aimed to strike a balance between basic and detailed annotation to accommodate an array of future analyses. We have defined various labels for each feature or component and organized these labels into macro categories to establish a coarser annotation level. This coarser level is expected to facilitate SLP tasks and statistical linguistic analyses.

Table 1 presents the main blocks of features covered by our annotation scheme, with each block corresponding to a set of tiers within an iLex transcript. In the following sections, we provide detailed explanations of the tiers included in each main block.

Video
Item / Task
Sentence
Manual components
Non-manual components
Errors
Additional information
Comments

Table 1: Main blocks of tiers in the transcription and annotation scheme.

### 4.1. Task Level

The initial segmentation of the video stream involves automatically annotating the task starting and ending times, along with the task code, in the **Item** tier.

Following this, each task time span is segmented into sentence-like units, which are labeled within the **Sentence** tier. These units may encompass anywhere from one to  $n$  sentences.

The segmentation process is subsequently extended to manual and non-manual components within each sentence.

### 4.2. Manual Components

In general, the most basic level of corpus annotation is tokenization. Tokens pertaining to manual components are identified and segmented within the sentence adhering to a wider segmenting system (Hanke et al., 2012).

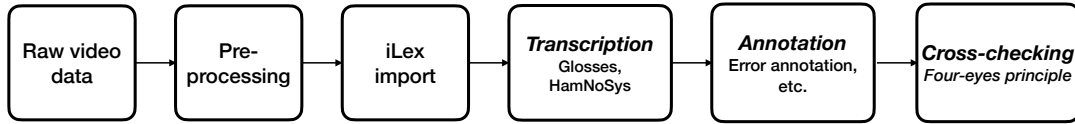


Figure 1: Visualization of the data process from raw data to data annotation.

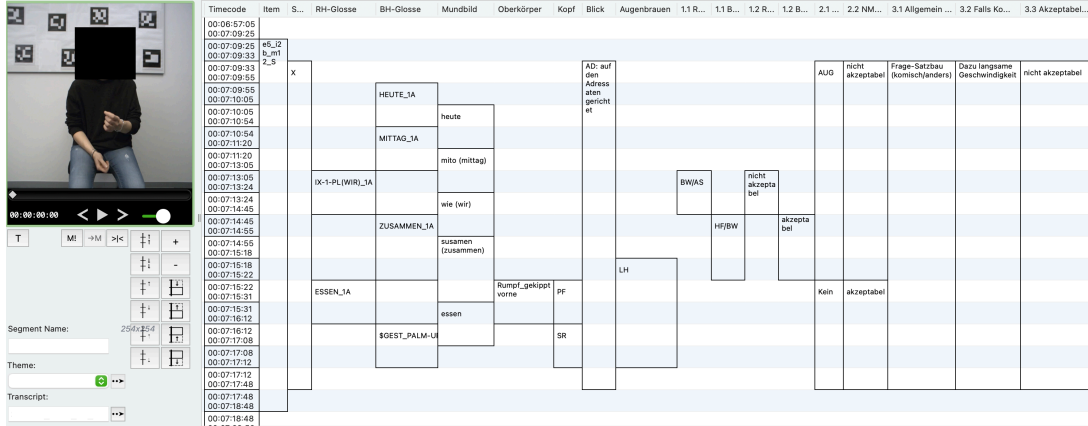


Figure 2: Sample transcript in iLex with manual, non-manual, and error annotation tier.

Table 2 outlines the tiers for the manual components included in our scheme. Following identification, manual components are annotated by inserting **identificative glosses** (ID glosses) as semantic notations, and described in their form using the **Hamburg Notation System for Sign Languages** (HamNoSys; Prillwitz, 1989). In using the iLex corpus lexicon system, we are assured of having consistent use of glosses by different annotators. The selection of glosses was motivated by their widespread usage as common semantic labels of signs. In addition, glosses are extensively employed in SLP, particularly in the domain of Sign Language Translation (SLT) (Müller et al., 2023).

In this phase, we distinguish between signs produced with the left or right hand as well as between one-handed and two-handed signs. The tier **Gloss Right Hand (RH)** is annotated for one-handed signs articulated on the right hand, while **Gloss Left Hand (LH)** is annotated for one-handed signs articulated on the left hand. Two-handed signs are annotated in **Gloss Both Hands (BH)**. The hand dominance of the signer is stored in the signer’s metadata.

Non-conventionalized signs, like gestures, are annotated similarly to glosses and allocated to the tiers of the hand used for articulation, identified by the affix *GEST\_*. Fingerspelling follows the same approach as single signs, annotated with the affix *FA\_* to the gloss.

Qualifiers are combined with glosses to indicate variant forms, involving slight differences in the phonological parameters (Konrad et al., 2012). The form variance is reported in the corresponding **Ham-**

**NoSys variance** tier. For glossing and qualifier addition, we adhere to the glossing conventions<sup>1</sup> of our iLex DSGS instance and those described in Konrad et al. (2012) and Ribeaud and Cicala (2019).

### Manual Components

#### Gloss RH

HamNoSys RH

HamNoSys variance RH

#### Gloss LH

HamNoSys LH

HamNoSys variance LH

#### Gloss BH

HamNoSys BH

HamNoSys variance BH

Table 2: Tiers of the manual components. RH: right hand. LH: left hand. BH: both hands.

### 4.3. Non-manual Components

Non-manual activities undergo detailed annotation in our scheme. Labels for each feature were based on the scheme for non-manual components in Hanke (2001), then determined on the most frequently annotated forms in previous DSGS studies and compared with those in studies outlined in Section 2. Each label specifies the form, movement, or both of a specific facial or body part compared to a neutral position. All labels were assigned an identifying code and accompanied by an image or

<sup>1</sup><https://dsgs-handbuch.ch/information/>

illustration available in iLex to facilitate the annotation process. At this stage, assignments of these labels to grammatical functions were not made. The complete annotation scheme for non-manual components is available in both German and English on Zenodo.<sup>2</sup>

Table 3 displays the tiers included in the non-manual components block of the annotation scheme. The **Mouthing** tier captures lip movements like those of spoken German words. As mouthings are often not exact pronunciations of words, the annotator inserts the letters representing what they observe during the lip movement of the signer displayed in the video. For example, in Figure 2, we can see how the mouthing “mito” was written for the word *Mittag* (‘noon’) because the final voiced velar consonant *g* does not involve any lip movement.

For **Mouth gestures**, annotators have the option to select from 81 labels. This is the most detailed part in our scheme, reflecting various nuances in the form and movement of mouth components such as lips, cheeks, teeth, tongue, and their combinations. These labels are grouped into nine macro categories based on the form rather than function of the labels, as was done for the Auslan corpus (Johnston, 2019).

Regarding the **Nose**, seven labels are defined and categorized as static or dynamic based on nose movement characteristics, such as *static wrinkled nose*.

In the **Upper body** tier, thirteen labels describe main movements, such as leaning or moving the torso in a specific way and subtly turning or rotating the torso so that it faces a particular direction. The direction is annotated from the signer’s point of view. **Shoulders** can be annotated separately from the upper body when their movements seem crucial to be considered in isolation, featuring six labels grouped under the macro categories of the upper body.

Fundamental in defining the sentence function, **Head** movements are segmented into twenty labels, subdivided based on movement type or location. Table 9 in Appendix B provides the list of head component labels.

Eye-related movements, namely **Eye gaze**, **Eye-brow** movements, and **Eyelid** motion, are segmented and labeled separately. In most of the tasks, the participant gaze is straight on the camera (cf. tier “Blick” (‘gaze’) in Figure 2). The annotation of gaze direction is crucial for marking the position or differences in object location. Eight labels denote various eyebrow positions, mostly upwards or downwards, while ten eyelid labels distinguish eye aperture and motion.

<sup>2</sup><https://doi.org/10.5281/zenodo.10669639>

Non-manual components
Mouthing
Mouth gesture
Nose
Upper body
Shoulders
Head
Eye gaze
Eyelids
Eyebrows

Table 3: Tiers of the non-manual components.

#### 4.4. Error Annotation

The error annotation tiers aim to capture productions by DSGS learners that deviate from the canonical form (Table 4). They are divided into three main categories: manual components, non-manual components, and sentence level.

For **manual components**, we adopted error definitions and categories from Ebling et al. (2018). These tiers, connected to gloss tiers, annotate deviations related to phonological parameters and their combinations.

For **non-manual components**, deviations regarding eyebrow and head movements, mouthing, mouth gestures, and their combinations are annotated. These features play a crucial role in sentence function definition.

The third category addresses **sentence-level error definition**. Drawing from prior studies and our main annotators’ long teaching experience, we defined a restricted list of error categories to start from: sentence construction, question construction, negation, affirmation, statement connection, indexing, verbs, signing space, tempo and fluency, combined issues, and others. Where one of the latter two categories is chosen, the annotators describe the corresponding errors in a free-text field of a separate tier.

Each deviation receives a degree of **(non)-acceptability** (*not acceptable*, *acceptable*, *fully acceptable*), indicating severity of the deviating feature and impact on sentence comprehension. Additionally, the entire sentence receives an acceptability value, regardless of the number of annotated deviations. Figure 3 illustrates a simplified annotation example of a sentence deemed as “not acceptable” due to incorrect sentence construction, such as the use of the mouthing “da” (‘there’) and the improper use of eyebrows in the sentence.

The *acceptability of the sentence* tier is also annotated for L1 data. The rationale behind this decision is explained in the next section.

Error annotation
Deviations Gloss RH
Acceptability
Deviations Gloss LH
Acceptability
Deviations Gloss BH
Acceptability
Deviations NMC
Acceptability
Sentence problem
Sentence acceptability

Table 4: Tiers of error annotation. NMC: non-manual components.

```

Head:      right |      | shaking      ||
Eye brows:      | furrowed |      ||
Mouthing: da | frau |      | kei | auto ||
Glosses:  IX-3 | FRAU | KEIN_bew | KEIN | AUTO ||
          |      | woman      | not | car
DE: Die Frau hat kein Auto.
EN: That woman does not have a car.

```

Figure 3: Example of the annotation of a “not acceptable” sentence.

#### 4.4.1. Why Annotate Acceptability?

Assuming a single “ground truth” in spoken and sign languages poses inherent challenges in achieving high agreement on language interpretation and understanding (Plank, 2022). Variations in annotation may arise from linguistic complexities, subjectivity, or instances where multiple interpretations are plausible (Plank et al., 2014; Manning, 2011; Rottger et al., 2022; Basile et al., 2021; Pavlick and Kwiatkowski, 2019; Nie et al., 2020). Sign languages are known to exhibit considerable structural variability (Bayley et al., 2015).

In the absence of a definitive ground truth, specifying acceptability values becomes more meaningful than assigning binary correct/incorrect values (Mehta and Srikumar, 2023). In the context of sign languages, the concept of acceptability of intuitive judgments was explored by Arendsen (2009) for the manual/phonological components of single signs in relation with iconicity. We thus designate sentences within an acceptable range from L1 data as correct, establishing them as the ground truth. Therefore, annotations of components in these acceptable sentences serve as a form of gold standard.

Having said this, we recognize that the annotation of acceptability values, like in error annotation, inherently entails a certain degree of subjectivity.

## 4.5. Additional Information

The additional tiers listed in Table 5 have not yet been systematically annotated at the current stage. This block of tiers is reserved for future rounds of annotations following preliminary linguistic analysis. In the interim, annotators may include comments in the *Comments* tier or annotate straightforward features. The **Translation** tier involves inserting a literal translation in German of individual signs and sentences. The **Functions**, **Topic/Focus**, **Prosody**, and **Role** tiers are designed to label various functions of annotated components, not only at the sentence level but also at the discourse level.

Additional information
Translation
Comments
Functions
Topic/Focus
Prosody
Role

Table 5: Additional tiers.

## 5. Validating the Annotation

As discussed in Section 3, our data undergoes a cross-checking step in which part of it is double-annotated. This step allows for the calculation of inter-annotator agreement (IAA) between the two expert annotators (Section 3), to assess the consistency of the (error) annotation labels, and to provide a quantitative evaluation of the complexity of the annotation task.

It is essential to recognize that agreement between annotators should not be mistaken with accuracy, as annotators may share possible biases present in the guidelines or cultural preconceptions (Basile et al., 2021; Plank, 2022).

### 5.1. Method

Incorporating different agreement metrics enabled us a thorough evaluation, considering various facets of annotation agreement. Applying Gwet’s *AC1* was motivated by specific limitations of Cohen’s  $\kappa$  (Cohen, 1960), particularly its tendency to underestimate coefficients for high-chance agreements and its lack of robustness against imbalanced categories (Feinstein and Cicchetti, 1990; Gwet, 2014).

In L1 data, we randomly extracted and duplicated 20% of the dataset, amounting to two transcripts. Each expert annotator annotated the transcript assigned to them and the counterpart annotated by



the other expert. We then extracted the annotations from iLex and computed agreement using the following methods. First, in each transcript sentence, we examined annotated time spans sharing the same feature annotation, computed the overlap proportion of each feature and then calculated the temporally weighted overlap ratio, as described in Andersson and Sandgren (2016). We reported the formula for calculating the ratio along with the explanation and an example in Appendix C. As illustrated in Figure 4 in Appendix C, we treated all labels within the same feature as identical.

Second, we calculated Cohen’s  $\kappa$ , Krippendorff’s  $\alpha$  (Krippendorff, 2019), and Gwet AC1 score for nominal data across all labels for each transcript. This analysis utilized macro categories for each annotated component, disregarding the time variable.

For L2 data, we randomly selected 20% of the annotated L2 sentences for the first two data collection points, amounting to a set of 38 sentences. Within these selected sentences, we introduced new tiers for error annotation while deactivating the original error annotation tiers. The second annotator reviewed the annotation of manual and non-manual components performed by the first annotator in the first and second rounds, and then carried out a new error annotation using only their initialized tiers. We then extracted the annotations from iLex and assessed reliability using Cohen’s  $\kappa$ , Krippendorff’s  $\alpha$ , and Gwet AC1 for nominal data. Agreement concerning acceptability values was evaluated using Cohen’s  $\kappa$ , Krippendorff’s  $\alpha$ , and Gwet AC2 score for ordinal data.

For error annotation of non-manual components, adjustments to the time span were made depending on the alleged occurrence of a non-manual component. Thus, we computed the overlap ratio and temporally weighted overlap ratio for this category, as outlined in Appendix C. For glosses and sentence-level annotation, we focused solely on the annotation label without considering timing. This choice stemmed from the consistent timing across annotators, established through prior segmentation and linkage of tiers in iLex.

## 5.2. Results

We acknowledge that direct comparison of the results from these methods is not feasible due to their differences in computation. Nevertheless, this initial exploration represents our first step toward a comprehensive evaluation of our annotated data.

Below, we present our preliminary findings regarding the validation of the data.

### 5.2.1. L1 Data

On average, the annotation of manual and non-manual components in the L1 data achieved an

overlap ratio of 0.18, encompassing cases for which the overlap duration is equal to 0. In instances of zero overlap, distinguishing missed events from misalignments was challenging. By excluding these events, the average overlap ratio increased to 0.62. Specifically, manual components attained an average of 0.64 (median: 0.88), while non-manual components averaged 0.45, ranging from 0.01 to 0.97. We calculate the temporally weighted overlap ratio for the events in each sentence. The average is 0.52, ranging from 0.29 to 0.96.

The agreement on labels is detailed in Table 6. Overall, the agreement between the two expert annotators did not reach high values. Considering both manual and non-manual components and excluding rows with zero overlap in time, the agreement yielded a  $\kappa$  score of 0.49 and a Gwet score of 0.52. Krippendorff’s values closely align with the  $\kappa$  scores.

	$\kappa$	$\alpha$	Gwet
manual	0.57	0.57	0.61
nmc	0.39	0.38	0.47
manual+nmc	0.49	0.44	0.52

Table 6: Reliability as measured by inter-annotator agreement using  $\kappa$ ,  $\alpha$ , Gwet AC1.

### 5.2.2. L2 Data

On average, the error annotation in the non-manual components of the L2 data achieved an overlap ratio of 0.35, ranging from 0.0 to 1 (median: 0.19). After excluding cases with zero overlap, the ratio increased to 0.55, ranging from 0.03 to 1 (median: 0.50). We calculated the temporally weighted overlap ratio for the events in each sentence obtaining an averaged score of 0.66.

Regarding the assigned labels, as presented in Table 7, agreement between the two expert annotators is modest.  $\kappa$  scores range from 0.16 for the error annotation of non-manual components to 0.52 for the error annotation of manual components, indicating a considerable degree of subjectivity in both annotation tasks. Krippendorff’s values closely mirror the  $\kappa$  scores.

Interestingly, the acceptability values for the error annotation of non-manual components achieved a Gwet score of 0.60, suggesting moderate to high agreement between the two expert annotators in assessing the severity of deviation for non-manual features.

## 5.3. Discussion

The level of agreement depends on the task, complexity of the annotation scheme, and the number of annotators along with their degree of expertise.

	$\kappa$	$\alpha$	<i>Gwet</i>
manual	0.52	0.53	0.56
accept_manual	0.32	0.33	0.34
nmc	0.16	0.15	0.25
accept_nmc	0.25	0.24	0.60

Table 7: Reliability as measured by inter-annotator agreement using  $\kappa$ ,  $\alpha$ , Gwet *AC1* (for components) or *AC2* (for acceptability). *Manual*: error annotation of the manual components; *nmc*: error annotation of the non-manual components; *accept*: agreement on the acceptability judgments.

Examining our results, the scores derived from our preliminary agreement calculations lead us to reflect on the primary factors contributing to disagreements.

Firstly, our findings underscore the inherent difficulty in achieving high agreement in tasks involving video stream segmentation. The accurate segmentation of signs presents challenges even for trained annotators, resulting in slight time variations in sign segmentation. However, these variations can cause discrepancies in calculations. In addition, the detailed nature of our annotation scheme, as described in Section 4, inherently amplifies disagreement among annotators. In general, studies analyzing sign language datasets refrain from reporting agreement scores, complicating efforts to benchmark our results within the broader landscape of sign language reliability assessments. The discrepancy between manual and non-manual component values (cf. Table 6 and Table 7) underscores the heightened challenge associated with annotating non-manual activities, possibly deriving from ambiguous guidelines or unclear instances of non-manual activity in videos.

Secondly, the complexity of the annotation task is reflected in the complexity of calculating agreement between annotators. Following the method outlined by Andersson and Sandgren (2016), which involves calculating the temporally weighted overlap ratio only between events with the same label, we do not assess whether there might be other annotated events occurring simultaneously but labeled differently. For instance, in cases where Annotator A annotated a time span with a label from the list of the “Eyelid” feature while Annotator B annotated the same time span with a label from the “Eyebrow” list, this could mean missing an event by one or both annotators. Considering the simultaneity of components in sign language, it is plausible that the time span involves both “Eyelid” and “Eyebrow” movements simultaneously. A next step would be to examine these “alternative classifications” with an aim to agree on one way of annotating and analyzing them.

As suggested by Schembri and Crasborn (2010),

further exploration into agreement calculations for sign language data is needed. Establishing annotation standards would facilitate comparison of agreement values across different corpora, allowing for the development of a systematic method for calculating agreement in sign language data.

Despite the relatively modest agreement values, it is imperative not to perceive them as a limitation for dataset validation and subsequent use of these annotations. Widely debated in the context of spoken languages, human label variation (in other words, disagreement) offers valuable data insights to consider in the development of technologies, particularly those aimed at enhancing “technology which is by and for humans; inclusive and reliable” (Plank, 2022).

## 6. Conclusion and Outlook

We have presented the annotation process and scheme for L1 and L2 DSGS continuous data, focusing on the labeling of non-manual components. We have introduced a method for annotating and categorizing linguistic errors in L2 data, and proposed our idea of creating a ground truth encompassing variability. Viewing sentence acceptability as a facet of ground truth expands traditional notions, accommodating the inherent variability in sign language data analysis.

Our annotation scheme remains a work in progress, open to modification and adaptation. Statistical analyses are warranted to evaluate the scheme’s efficacy and the utility of macro categories. Refinement on higher levels of annotation, such as on the levels of sentence function and semantic roles (some tiers are described in Section 4.5), remains an area for future development.

While the scheme was created for DSGS, it can be adapted to other sign languages by adjusting the labels of each feature. To maintain the application of cross-linguistic comparisons, the adjustments would not change the content of the components but only the names that are assigned to these components.

We have described our first approach to data validation, illustrating difficulties given by the different variables to consider in the calculation. Agreement calculation methods, particularly considering time spans and labels, demand further exploration to systematically analyze annotated events and spot missed or erroneously annotated instances.

As we move forward, collaborative efforts and continued refinement of annotation practices will facilitate the advancement of sign language research.

## Acknowledgments

This work was funded through the Swiss National Science Foundation (SNSF) Sinergia project “Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment II” (SMILE II) (grant agreement no. CRSII5\_193686).

The authors would like to thank Regula Perrollaz, and Annemarie Büchli-Meier for their annotation work, and Penny Boyes Braem for her insightful assistance, comments, and reviews.

## 7. Bibliographical References

- Richard Andersson and Olof Sandgren. 2016. [ELAN Analysis Companion \(EAC\): A Software Tool for Time-course Analysis of ELAN-annotated Data](#). *Journal of Eye Movement Research*, 9(3).
- Jeroen Arendsen. 2009. *Seeing signs: on the appearance of manual movements in gestures*. publisher not identified, Place of publication not identified. OCLC: 823212702.
- Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. [Toward a perspectivist turn in ground truthing for predictive computing](#). *CoRR*, abs/2109.04270.
- Robert Bayley, Adam C. Schembri, and Ceil Lucas. 2015. *Variation and change in sign languages*, page 61–94. Cambridge University Press.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Brafport, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Kearsey Cormier, Onno Crasborn, and Richard Bank. 2016. [Digging into Signs: Emerging Annotation Standards for Sign Language Corpora](#). In *7th Workshop on Representation and Processing of Sign Languages: Corpus Mining*, pages 35–40, Portorož, Slovenia. ELRA.
- Sarah Ebling, Necati Cihan Camgoz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi, and Mathew Magimai-Doss. 2018. [SMILE Swiss German sign language dataset](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4221–4229, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sarah Ebling, Katja Tissi, Sandra Sidler-Miserez, Cheryl Schlumpf, and Penny Boyes Braem. 2021. [Single-parameter and parameter combination errors in L2 productions of Swiss German Sign Language](#). *Sign Language & Linguistics*, 24(2):143–181.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. [High agreement but low Kappa: I. the problems of two paradoxes](#). *Journal of Clinical Epidemiology*, 43(6):543–549.
- Silvia Gabarró-López and Laurence Meurant. 2014. [When nonmanuals meet semantics and syntax: a practical guide for the segmentation of sign language discourse](#). In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavik, Iceland.
- Gaëtanelle Gilquin and Sylvie De Cock. 2011. [Errors and disfluencies in spoken corpora: Setting the scene](#). *International Journal of Corpus Linguistics*, 16(2):141–172. Publisher: John Benjamins Type: Journal Article.
- Sannah Gulamani, Chloë Marshall, and Gary Morgan. 2020. [The challenges of viewpoint-taking when learning a sign language: Data from the ‘frog story’ in British Sign Language](#). *Second Language Research*, 38(1):55–87.
- Kilem Li Gwet. 2014. *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*, fourth edition edition. Advances Analytics, LLC, Gaithersburg, Md.
- Thomas Hanke. 2001. [Visicast deliverable d5–1: Interface definitions](#). technical report. visicast project. Technical report.
- Thomas Hanke, Silke Matthes, Anja Regen, and Satu Wörseck. 2012. [Where Does a Sign Start and End? Segmentation of Continuous Signing](#). In *5th Workshop of the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon (LREC 2012)*., Istanbul, Turkey. ELRA.
- Thomas Hanke and Jakob Storz. 2008. [iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography](#). In *Proceedings of the 6th Language Resources*



- and Evaluation Conference (LREC), pages 64–67, Marrakesh, Morocco. ELRA.
- Gabrielle Hodge. 2014. *Patterns from a signed language corpus: Clause-like units in Auslan (Australian sign language)*. Ph.D. thesis, Macquarie University.
- Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In Trevor Johnston, Julie A. Hochgesang, and Jordan Fenlon, editors, *Signed Language Corpora*, pages 46–89. Gallaudet University Press, United States.
- Trevor Johnston. 2010. *From archive to corpus: Transcription and annotation in the creation of signed language corpora*. *International Journal of Corpus Linguistics*, 15(1):106–131. Publisher: John Benjamins Publishing Company.
- Trevor Johnston. 2019. *Auslan Corpus Annotation Guidelines*.
- Reiner Konrad. 2011. *Die lexikalische Struktur der Deutschen Gebärdensprache im Spiegel empirischer Fachgebärdenlexikographie*. Ph.D. thesis, Universität Hamburg.
- Reiner Konrad, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio, and Anja Regen. 2012. *From form to function. a database approach to handle lexicon building and spotting token forms in sign languages*. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 87–94, Istanbul, Turkey. European Language Resources Association (ELRA).
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. *Public DGS Corpus: Annotation Conventions / Öffentliches DGS-Korpus: Annotationskonventionen*.
- Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022. *Specification for the Harmonization of Sign Language Annotations*.
- Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition edition. Thousand Oaks, California.
- Kim B. Kurz, Geo Kartheiser, and Peter C. Hauser. 2023. *Second language learning of depiction in a different modality: The case of sign language acquisition*. *Frontiers in Communication*, 7.
- Andrea Lackner. 2019. *Describing Nonmanuals in Sign Language*. In Andrea Lackner, editor, *Grazer Linguistische Studien*, volume 91, pages 45–103. University of Graz, Graz, Austria.
- Christopher D. Manning. 2011. *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?* In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6608, pages 171–189. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Maitrey Mehta and Vivek Srikumar. 2023. *Verifying annotation agreement without multiple experts: A case study with Gujarati SNACS*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.
- Johanna Mesch, Krister Schönström, Nikolaus Riemer Kankkonen, and Lars Wallin. 2016. *The interaction between mouth actions and signs in swedish sign language as an l2*. In *Presented at the The 12th International Conference on Theoretical Issues in Sign Language Research (TISLR)*.
- Johanna Mesch and Krister Schönström. 2018. *From Design and Collection to Annotation of a Learner Corpus of Sign Language*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Johanna Mesch and Krister Schönström. 2020. *Use and acquisition of mouth actions in L2 sign language learners: A corpus-based approach*. *Sign Language & Linguistics*, 24(1):36–62. Publisher: John Benjamins Publishing Company.
- Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. *Considerations for meaningful sign language machine translation based on glosses*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.
- Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. *What can we learn from collective human opinions on natural language inference data?* In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.
- Annika Nonhebel, Onno A. Crasborn, and Els van der Kooij. 2004. *Sign language transcription conventions for the echo project (version 9)*. Technical report.
- Gerardo Ortega and Gary Morgan. 2015. *Phonological Development in Hearing Learners of a Sign*



- Language: The Influence of Phonological Parameters, Sign Complexity, and Iconicity: Phonological Development in Sign L2 Learners. *Language Learning*, 65(3):660–688.
- Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.
- Siegmund Prillwitz. 1989. *HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide*. Intern. Arb. z. Gebärdensprache u. Kommunik. Signum Press.
- Marina Ribeaud and Isabelle Cicala. 2019. *Handbuch Glossierung der Deutschweizerischen Gebärdensprache (DSGS)*, second edition. fingershop.ch.
- Russel S. Rosen. 2004. Beginning L2 production errors in ASL lexical phonology: A cognitive phonology model. *Sign Language & Linguistics*, 7(1):31–61.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.
- Adam Schembri and Onno Crasborn. 2010. Issues in creating annotation standards for sign language description. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valletta, Malta.
- Marc Schulder, Sam Bigeard, Thomas Hanke, and Maria Kopf. 2023. The sign language interchange format: Harmonising sign language datasets for computational processing. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.
- Krister Schönström. 2021. Sign languages and second language acquisition research: An introduction. *Journal of the European Second Language Association*, 5(1):30–43.
- Krister Schönström and Johanna Mesch. 2014. Use of nonmanuals in adult L2 signers in Swedish Sign Language – Annotating the nonmanuals. In *6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, Reykjavik, Iceland. ELRA.
- Katja Tissi. 2021. *DSGS-Handbuch. Interkantonale Hochschule für Heilpädagogik Zürich HfH*.
- Lars Wallin and Johanna Mesch. 2018. *Annoteringskonventioner för teckenspråkstexter : Version 7 (januari 2018)*. Technical report, Stockholm University, Sign Language.
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).

## 8. Language Resource References

- Crasborn, Onno and Zwitserlood, Inge and Ros, Johan and van Kampen, Annemieke. 2006-2017. *Collection Corpus NGT*. The Language Archive. PID <https://hdl.handle.net/1839/8e5a77a3-8d1a-492a-bc86-9a3398b0809c>.
- Trevor Johnston. 2008. *Auslan Corpus*. Endangered Languages Archive. PID <http://hdl.handle.net/2196/00-0000-0000-0000-D7CF-8>.

### A. Non-manual Components: Macro categories

#### B. Head Labels

The two-letter codes, an extension of HamNoSys to non-manuals, were initially defined in the ViSiCAST project (Hanke, 2001) and have been adapted by our annotators.

Feature	Labels	Macro categories
Mouthing	-	-
Mouth gesture	81	9
Nose	7	2
Upper body	13	2
Shoulder	6	1
Head	20	6
Eye gaze	30	6
Eyelids	10	3
Eyebrows	8	2

Table 8: Number of labels and number of macro categories in our scheme.

Head	Macro categories
NO: Head nod (up and down) NU: Simple head nod up [dynamic] ND: Simple downward head nod [dynamic] RL: Tilted to left or right nodding head	cat. 1 Nodding
SH: Head shaking (left and right) SS: Tilted to left or right shaking head	cat. 2 Shaking
NF: Tilted forward [static] PF: Shifted forward OG: Head tilted forward (nodding)	cat. 3 Front
NB: Tilted backwards PB: Shifted backward LN: Head nod (up and down) left (up and down) RN: Head nod (up and down) right (up and down)	cat. 4 Back
SL: Turned to the left SR: Turned to the right TL: Tilted to the left (static) TR: Tilted to the right (static)	cat. 5 Lateral
KD: Head rotation KK: Head tilt (dynamic) LI: Head movement coupled to gaze [dynamic]	cat. 6 Strongly dynamic

Table 9: Labels defined for the Head feature.

### C. Temporally weighted overlap ratio

Equation 1 illustrates an example of the agreement calculation with two events in the L1 data, as illustrated in Figure 4. Column A represents two events for the feature “Blick” (‘gaze’) annotated by Annotator A in one sentence, while Column B represents the two events in the same sentence annotated by Annotator B. We have:

$$\begin{aligned}
 E &= \{\epsilon_1, \epsilon_2\} \\
 T &= \{t_1 = 0.39, t_2 = 2.76\} \\
 O &= \{o_1 = 0.05, o_2 = 0.95\}
 \end{aligned} \tag{1}$$

where  $E$  is the set of  $n = 2$  events, each labeled by Annotator A and Annotator B;  $T$  is the set of maximum duration for each events in  $E$ , and  $O$  represents the set of overlap proportions for the events in  $E$ . The overlap proportion is calculated by dividing the duration of the overlap by the maximum temporal extent of the event.

The temporally weighted overlap ratio is then calculated as follow:

$$\frac{\sum_i^n O_i T_i}{\sum_i^n T_i} = \frac{(0.05 * 0.39) + (0.95 * 2.76)}{(0.39 + 2.76)} = 0.84 \tag{2}$$

If we were to consider only the overlap proportion without accounting for temporal duration, the calculation for the overlap ratio would be as follows:  $0.05 + 0.95/2 = 0.5$ , even though the length of the annotated overlap varies.

Time span	A	B
00:08:33.06 00:08:33.35	Blick 1	
00:08:33.35 00:08:33.37		Blick 3
00:08:33.37 00:08:33.45		
00:08:33.45 00:08:35.53	Blick 1	Blick 1
00:08:35.53 00:08:36.11		
00:08:36.11 00:08:36.21		

Figure 4: Simplified representation of two events in a same sentence, annotated by two annotators, Annotator A and Annotator B.

Please note that even if the annotators assigned two different labels for event  $\epsilon_1$ , they both annotated the feature “Blick” (‘gaze’) in this timespan.

# Person Identification from Pose Estimates in Sign Language

Alessia Battisti\* , Emma van den Bold\* , Anne Göhring\* ,  
Franz Holzknicht† , Sarah Ebling\* 

\*University of Zurich

Andreasstrasse 15, 8050 Zurich

{battis, goehring, ebling}@cl.uzh.ch | emma.vdbold@gmail.com

†University of Teacher Education in Special Needs

Schaffhauserstrasse 239, 8050 Zurich

franz.holzknicht@hfh.ch

## Abstract

Sign language recognition models require extensive training data. Effectively anonymizing such data remains a complex endeavor due to the crucial role of facial features. While pose estimation techniques have traditionally been considered a means of yielding anonymized data, the findings reported in this paper challenge this assumption: We conducted a study involving Swiss German Sign Language (DSGS) users, presenting them with pose estimates from DSGS video samples. The participants' task was to identify the signers' language levels and identities from skeletal representations. Our findings reveal that the extent to which sign language users were capable of recognizing familiar signers depended on their language level, with deaf experts achieving the highest accuracy. We demonstrate that an automatic classifier obtains comparable results in multi-label language level recognition ( $F1=0.64$ ) and person identification ( $F1=0.31$ ). This emphasizes the need to reconsider the fundamentals of video anonymization towards guaranteeing sign language users' privacy.

**Keywords:** Data anonymization, sign language videos, pose estimation

## 1. Introduction

In recent years, more and more studies have been published in the area of automatic sign language processing (SLP), including Sign Language Translation (SLT) (Bull et al., 2020; De Sisto et al., 2021; Varol et al., 2021; Momeni et al., 2022; Müller et al., 2022, 2023). The growth of this field has intensified the demand for sign language data, opening a discussion about the privacy of sign language users who share their data in research (Bragg et al., 2020) and on social media platforms (Mack et al., 2020).

The topic of anonymization of sign language data has thus become relevant in several areas of research, from the improvement of accessible design to the enhancement of SLP for new technologies (Bragg et al., 2020; Lee et al., 2021; Xia et al., 2022, 2023). The collection and use of sign language data is challenging due to privacy concerns and ethical considerations (Bragg et al., 2020). Sign language users may feel uncomfortable participating in research and sharing data due to a lack of video anonymization methods that protect their privacy.

Enhanced privacy could lead to an increased participation of sign language users in research and to an improvement of SLP results (Bragg et al., 2021). The development of effective anonymization techniques is therefore a necessary precursor.

Anonymizing sign language data is not a trivial task due to the visual-gestural nature of the language and the lack of a common writing system.

Obscuring or masking non-manual components, e.g., in the face would severely compromise the meaning and, consequently, the comprehension of utterances.

The SLP field widely uses pose estimation systems that generate skeleton-like representations from persons in videos (Stoll et al., 2020; Saunders et al., 2021, 2022). As such, there has been an increasing perception that pose estimation systems can be employed for anonymizing sign language data. Whether the skeleton-like representations do, in fact, sufficiently conceal the identity of the signers underlying the pose estimates is an open question.

Given this context, we conducted an online visual perception study for Swiss German Sign Language (*Deutschschweizerische Gebärdensprache*, DSGS) and investigated whether sign language users were able to correctly identify the language level (RQ1) and the identity (RQ2) of the signers displayed in short videos processed with pose estimation technology. We hypothesized that signers with different levels of DSGS could identify signers to a different extent. We additionally assessed the participants' comprehension of the linguistic content of sentences represented in skeletal form and we used this information to train two classifiers to assess the automation of the tasks of language level recognition and person identification. Finally, we looked for patterns in the factors that led to correct identification in each group.

It is worth mentioning that the DSGS community

is relatively small, as is the case of many deaf<sup>1</sup> communities around the world. There are an estimated 5,500 native signers/early learners<sup>2</sup> of DSGS and an additional 13,000 hearing users with different connections to sign language, such as through education, social work, having a deaf family member, or just being interested in the language (Boyes Braem et al., 2012). Therefore, the chances of identification, as well as the potential consequences, can be considerable (Crasborn, 2008).

To the best of our knowledge, our study represents the first effort in addressing the identifiability of sign language users through pose estimates. This study is the first investigation to include DSGS users, paying unique attention to a low-resourced sign language. Lastly, the study provides pointers to future work in sign language data anonymization, highlighting important aspects to consider when anonymizing videos to guarantee privacy to sign language users.

## 2. Related Work

Existing computer vision algorithms used in pose estimation for SLP often ignore privacy concerns and rely on high-resolution image capture (Hinojosa et al., 2021). Privacy-preserving pose estimation typically involves reducing image resolution or distorting the image, sometimes combining multiple approaches (Jiang et al., 2022). However, these strategies are not suitable for sign language data, as they may compromise the linguistic content of the videos.

Similarly, early sign language anonymization techniques tended to compromise the linguistic content by modifying or hiding visual features of the individuals in the videos, which effectively prevent facial identification (Bleicken et al., 2016; Isard, 2020). Appendix A shows examples of blackening (Figure A.1a), blurring (Figure A.1b), and masking with filter (Figure A.1c).

In contrast, newer systems, based on generative neural networks, are capable of modifying signers' appearances and reproducing facial expressions while retaining the original linguistic content. Pose estimation techniques receive a sequence of raw images of a person as input and compute the positions and orientations of key body joints to generate skeleton-like representations of that person (Cao et al., 2021). In this way, information on the location

of various body parts is retained, while information on the appearance of the person and background is discarded. OpenPose<sup>3</sup> (Cao et al., 2019) was applied along with the above-mentioned blackening method to anonymize the data of the Public German Sign Language Corpus (Isard, 2020; Schulder and Hanke, 2020).

Recently, skeletal representations have been used to generate new images (Saunders et al., 2021; Xia et al., 2023) and avatars (Tze et al., 2022). Saunders et al. (2021) use pose estimates to eliminate the appearance of the input video, but retain motion information to reproduce the linguistic content of signed utterances (Figure A.1d). Their system then synthesizes a sequence of images of a signer with an appearance different from that of the input video. In Lee et al. (2021), the authors evaluate the effectiveness of various masking approaches and, consequently, their level of anonymization. They exploit a system that changes the identity of signers by replacing their face with the face of another person, maintaining linguistic information. Xia et al. (2022) extend this model towards full-body anonymization. They perform a similar process as in Saunders et al. (2021) but without leveraging pose estimation. The resulting model shows promising results, although preservation of linguistic content is not assessed.

Motion capture systems are capable of generating pose estimates as well (Gibet, 2018; Bigand, 2021). They utilize sensors to capture and replicate the motion of an individual's face and body, but their implementation is expensive and invasive due to the required equipment (Figure A.1e). These systems have found application primarily in the field of kinematic studies (Loula et al., 2005; Bigand et al., 2020). Within these investigations, it has been demonstrated that movement serves as a distinctive trait among individuals, facilitating their identification based on motion patterns. In the context of sign language motion studies, the work of Bigand et al. (2020) has shown that deaf observers are capable of recognizing signers based on motion capture data alone, emphasizing the need for techniques to conceal movement aspects. While Bigand et al.'s study focuses on identifying signers through motion capture data to explore how human traits are encoded in motion patterns, our study shifts the identification challenge to the domain of sign language research. Specifically, we target the recognition of poses generated by pose estimation techniques, by simulating a real-world scenario within a relatively small deaf community. Our primary focus is practical, addressing the current level of anonymity of pose estimates and assessing their limitations.

---

<sup>1</sup>We follow the recent convention of abandoning a distinction between "Deaf" and "deaf", using the latter term also to refer to (deaf) members of the sign language community (Napier and Leeson, 2016; Kusters et al., 2017).

<sup>2</sup>In this group, we include not only signers born to a deaf parent but also deaf signers who use DSGS as their primary language and acquired it at an early age.

---

<sup>3</sup><https://github.com/CMU-Perceptual-Computing-Lab/openpose>



### 3. Study Design and Data Collection

#### 3.1. Participants

In our study, we distinguished between two groups of participants: signers (**S**), who appeared in the study videos, and raters (**R**), who provided their responses as part of the online survey.

We were interested in investigating whether the language level affected person identification, therefore both signers and raters were grouped into three groups according to the language level: deaf native signers/early learners of DSGS (referred to **DE** for deaf expert), professional DSGS hearing interpreters with advanced language knowledge (**I** for interpreter), and hearing learners of DSGS with beginner skills (**L** for learners).

We recruited 21 raters by collaborating with research initiatives focused on DSGS at two Swiss universities. To participate in the study, IR and LR had to have knowledge of DSGS to the extent of at least level A1 (L group) and B2 (I group) according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2009) and be familiar with all or part of the signers in the videos used in the study (Section 3.3).

All signers and raters provided their informed consent, with the option to withdraw from the study at any time. Ratets were compensated in the form of either money or, for LR, course credits towards their studies.

Table 1 reports the total number of participants in the role of raters and signers for each language level group. Six raters appeared in the study stimuli themselves, i.e., they were also signers (Rater=Signer column). This overlap allowed us to investigate whether the signers were capable of identifying themselves.

Language Level	Raters	Signers	Rater=Signer
DE	4	3	2
I	4	3	1
L	13	3	3
Total	21	9	6

Table 1: Total number of raters and signers for each language group. The last column on the right shows the number of raters who also appeared as signers.

#### 3.2. Stimuli

We selected 45 videos from three existing datasets. For each signer, we manually selected five segments that were trimmed so as to adhere to linguistic content units. Each segment contained between 1 and 4 complete sentences (median: 2.0) and between 5 and 25 glosses (mean: 13.91) in a time span of 7 to 12 seconds (mean: 10.31  $\pm$  2.17).

Pose sequences were generated from the front view of the segments using MediaPipe Holistic (Grishchenko and Bazarevsky, 2020).<sup>4</sup> Figure A.1f in Appendix A displays an example of a pose produced from one sample.

#### 3.3. Survey

Raters were asked to watch the videos of the signers and answer a number of questions in the form of an online survey. They completed the survey on their laptops in a single session on the same day. Three key aspects were evaluated through a questionnaire combining qualitative and objective assessment methods. First, raters were tasked with assessing their **comprehension and fluency** of the sentences displayed as pose sequences, rating on a Likert scale ranging from 1 (*Not at all comprehensible/fluent*) to 4 (*Very comprehensible/fluent*). Additionally, raters were requested to transcribe utterances using DSGS glosses or translate them into German for an objective comprehension assessment. Second, the assessment focused on **language level identification**, presenting pose sequences categorized under three signer language levels, and offering options such as “deaf signer who knows DSGS well”, “hearing person who is an advanced user of DSGS”, and “hearing person who is a beginning learner of DSGS.” Last, the survey included questions related to **signer identification**, prompting raters to identify and name the signers depicted in skeletal representations, along with a brief justification based on the factors contributing to their identification.

To confirm whether the raters indeed knew all of the signers, we conducted a follow-up survey in which we showed them a video clip of each signer, as opposed to a pose sequence representing the signer.

### 4. Methods

Prior to explaining the methods, we present our research questions in detail:

**RQ1 Language level identification:** **RQ1.1** Are sign language users capable of identifying (other) signers’ language levels based on pose sequences? **RQ1.2** Where language level identification is successful, what are the factors that contribute to it? **RQ1.3** Can a classifier identify the language level using the same factors as sign language users?

**RQ2 Person identification:** **RQ2.1** Are sign language users capable of identifying signers that are *known to them* from pose sequences?

<sup>4</sup><https://github.com/J22Melody/pose-pipelines>

**RQ2.2** Where person identification is successful, what are the main factors that contribute to it? **RQ2.3** Can a classifier identify a signer using the same factors as sign language users?

#### 4.1. Calculating Identification Accuracy

The goal of **RQ1.1** was to assess the raters' ability to correctly determine the language level of the signers based on pose estimates. Therefore, we calculated the ratio of correct answers to the total number of answers within each signer group to measure identification accuracy for the language level.

In order to answer **RQ2.1**, we computed identification accuracy as the ratio of correctly identified signers to the total number of answers for each signer group. Additionally, we calculated accuracy at the individual signer level, i.e., by dividing the number of correct answers for each signer by the total number of answers related to that signer.

To address both **RQ1.2** and **RQ2.2**, we compared the raters' transcriptions of each content stimulus with the gold standard for that specific utterance, assuming that the comprehension of the linguistic content could potentially affect the capability of (correctly) determining the language level and identity of the signers. We hypothesized that higher similarity values could correspond to improved comprehension of the linguistic content of the stimuli, potentially enhancing the ability to identify the signer's language level and identity. For this, we calculated cosine similarity scores comparing the sentence embeddings (Reimers and Gurevych, 2019) of the transcriptions and the gold standards generated using a multilingual pre-trained language model, suitable for German<sup>5</sup>.

Finally, we examined the distribution of comprehension and fluency values assigned by the raters to each stimulus and related them to the identification accuracy.

#### 4.2. Designing Identification Classifiers

Using the collected data, we trained two multi-label support vector machine (SVM) classifiers: the first for the task of determining the language level between the three language categories ("language level classifier"; **RQ1.3**), and the second to discern signers ("signer classifier"; **RQ2.3**). We chose SVMs for explainability reasons.

The language classifier predicted the language level of the signers based on the raters' comprehension and fluency ratings as well as the number of glosses contained in the gold standard transcription

of the utterances. Including the latter feature was motivated by our hypothesis that a higher quantity of signs (as measured in glosses) produced by the signer within a given time frame imparts greater comprehension difficulty on the rater.

The signer classifier was trained to distinguish among the nine signers. As with the language classifier, it was based on comprehension and fluency ratings and the number of glosses in the utterances. As a baseline, we designed a dummy model that makes predictions based on the most frequent class label in the dataset, ignoring the input feature values.

We then employed 10-fold cross validation to test the performance of both classifiers, optimized through grid search. Considering only the comprehension and fluency features, we speculated that a deviation in performance between the classifiers and raters might suggest the presence of factors in human evaluation that were not explicitly collected through our survey and could not be reproduced by the classifiers.

#### 4.3. Annotating the Justifications

To further investigate the factors that contributed to successful identification of signers (**RQ2.2**), we analyzed the data collected using qualitative and quantitative methods. We performed an inductive qualitative coding (Skjott Linneberg and Korsgaard, 2019) to identify common themes (factors) relevant for the alleged identification of signers by the raters.

We used a collaborative process to code all free-text answers and create the codebook. After a first screening of all answers, we defined an initial set of codes that corresponded to the themes expressed explicitly or implicitly in the responses. Each answer was then allocated one or multiple codes, depending on the content. Three of the authors then iteratively refined and divided the list of codes into main themes and sub-themes, following fundamental concepts of sign language linguistics. The annotations were performed separately and then combined. Annotations that did not overlap were discussed among the annotators to arrive at a unanimous decision.

Overall, we labeled 195 answers; of these, 117 were based on correct identifications of signers.

The final codebook is shown in Appendix B. The anonymized dataset and annotated justifications are published on Zenodo.<sup>6</sup>

<sup>5</sup><https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1>

<sup>6</sup><https://doi.org/10.5281/zenodo.10669768>

## 5. Results

### 5.1. Quantifying Language Level Identification

To answer **RQ1.1**, we examined the responses pertaining to all rater-signer pairs (i.e., including cases where a rater had indicated not knowing a signer in our follow-up survey), assuming that it is possible to identify a signer’s language level even without being familiar with them. Table 2 reports the number of correct language level identifications and the corresponding accuracy across rater and signer groups. Different denominators resulted from different numbers of raters per group (Table 1). Overall, raters correctly identified the language levels 616 out of 934 times, resulting in a total accuracy of 65.95%. DERs achieved the highest accuracy (85%), with particular precision in identifying the ISs (91.67%). Among the signer groups, the learner language level was the most correctly identified across rater groups (85.48%).

### 5.2. Investigating Language Level Identification

#### 5.2.1. Factors Contributing to Identification

The distribution of correct and incorrect identifications against similarity values shows that higher similarity values correspond to accurate language level identifications, with variations among groups (Figure E.3 in Appendix E). For the DER group, average similarity scores remain consistent between correct and incorrect identifications (both around 0.7). In contrast, IRs and particularly LRs demonstrate a link between accurate identification of language levels and comprehension of the content, leading to more precise transcriptions.

Focusing only on correct answers, the LRs easily recognized the language levels of their peers and obtained higher similarity scores in the transcriptions of their utterances (Figure E.4 in Appendix E). This pattern could be attributed to learners’ tendencies to use simpler signs and sign at a slower pace, resulting in sentences that are easier to understand. A statistically significant correlation of 0.324 ( $p = 0.0$ ) between correct language level identifications and similarity scores is found exclusively for the LR group.

Examining only the comprehension aspect, we observed a decrease in comprehension ratings as rater language levels decline (Figure E.5 in Appendix E, left). DERs assigned higher comprehension scores, suggesting better subjective understanding, while LRs reported minimal comprehension. Regarding fluency, the ratings rise as signer language levels increase (Figure E.5 in Appendix E, right). LSs seldom achieve high fluency scores,

aligning with the perception that lower language level signers are perceived as less fluent. Especially, ISs received comparable high fluency ratings to DESs, suggesting interpreters were perceived as nearly as fluent as deaf experts.

#### 5.2.2. Automatic Classification of Language Levels

To answer **RQ1.3**, we explored the results of the multi-label language classifier reported in Table D.6 in Appendix D. Figure 1 shows the confusion matrix of the language classifier, over a 10-fold cross-validation on all data: While LSs were almost never confused, there is some overlap between DESs and ISs. Similarly, LRs made the same mistake by confusing DESs and ISs in the survey responses.

To deeper investigate this outcome, we designed a binary classifier for each language level to predict whether a signer had that specific language level (e.g., DE), based on the same predictive features of the language classifier. DESs were the most difficult category to be recognized, obtaining an F1 score of 0.55. Conversely, the LSs were the most correctly classified, with  $F1=0.85$ .

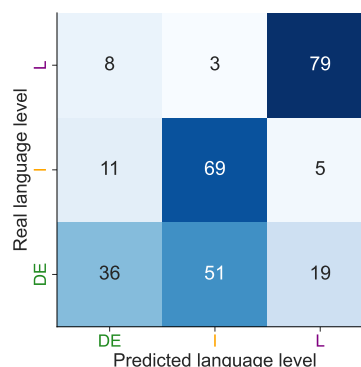


Figure 1: Confusion matrix for the language classifier predicting signers’ language levels, evaluated using 10-fold cross-validation.

The final classifier ‘DE+I+L’ obtained an F1 score of 0.638 and reached an accuracy of 65.7%, which is almost equivalent to the total accuracy of 65.95% obtained by the raters (Table D.6 vs. Table 2). In comparison, the dummy model obtained an F1 score of only 0.168.

### 5.3. Quantifying Person Identification

In addressing **RQ2.1**, the question on the correct identification of familiar signers, our analysis considered raters who knew the signers. All raters were familiar with all signers, except for one signer from the I group and two from the L group (Figure C.2 in Appendix C).

Groups	DES	IS	LS	Total
<b>DER</b>	<b>44/60 (73.33%)</b>	<b>55/60 (91.67%)</b>	<b>54/60 (90.0%)</b>	<b>153/180 (85.0%)</b>
<b>IR</b>	25/55 (45.45%)	45/59 (76.27%)	48/55 (87.27%)	118/169 (69.82%)
<b>LR</b>	102/195 (52.31%)	80/195 (41.03%)	163/195 (83.59%)	345/585 (58.97%)
<b>Total</b>	171/310 (55.16%)	180/314 (57.32%)	265/310 (85.48%)	<b>616/934 (65.95%)</b>

Table 2: Number of correct language identifications (percentages in brackets) across language groups. Values in bold indicate the highest scores for each signer group, and the total score.

Table 3 illustrates that raters achieved a total of 117 correct identifications, resulting in an overall accuracy of 13.64%. Accuracy exhibited a considerable dependence on signer and rater language levels. The better performance of the DERs compared to the other two groups could be potentially attributed to their more advanced receptive skills, a characteristic well studied in sign language linguistics, that improve along with the development of language proficiency (Beal-Alvarez, 2016; Hall and Reidies, 2021; Johnston, 2004).

Examining individual signers, Table 4 shows an even higher variability in accuracy. DERs consistently identified the three DESs correctly, with accuracy ranging between 35% and 45%. Signer 5, a well-known interpreter working for the Swiss national broadcaster, was correctly identified with an accuracy of 55% by DERs, 73.7% by IRs, but only 3% by LRs.

LSs had lower identification rates, with DERs achieving 80% accuracy for Signer 7. IRs never correctly identified any of the learners, potentially linked to lower familiarity.

Focusing on raters who also appeared as signers in the stimuli, five out of six identified themselves correctly in at least one instance. DERs achieved 80% accuracy, IRs 40%, and LRs 13%. This self-identification trend may be tied to receptive skill development and the ability to recognize one’s own movements, as supported by previous kinematics studies (Bigand et al., 2020; Loula et al., 2005).

## 5.4. Investigating Person Identification

### 5.4.1. Factors Contributing to Person Identification

To answer RQ2.2, we first investigated the distribution of correct identifications between signer groups based on similarity scores to determine whether a discernible pattern emerged (Figure F.6 in Appendix F). We found a weak positive Pearson correlation of 0.175 ( $p - value < 0.005$ ) between the similarity scores and the correct signer identifications. Comprehension as manifested through accurate transcription of the signed utterances did not influence the correct identification of signers. However, we observed a distinction between the similarity scores obtained in the transcription of

utterances produced in correct and incorrect identifications within the LRs, as already described for language level identification in Section 5.2. The transcriptions in which the signer was identified obtained a higher average similarity score compared to the transcriptions of the utterances where the signer was not correctly identified.

We investigated the comprehension and fluency ratings. As with the linguistic level identification task, for the signer identification task, we also noticed analogous rating distributions for comprehension. Both DERs and IRs never assigned the lowest comprehension score in conjunction with correctly identified signers (Figure F.7 in Appendix F, left).

With regard to fluency (Figure F.7 in Appendix F, right), the signer groups obtained high ratings, especially the interpreters. Among the correct responses, raters with higher language levels had a better understanding of the linguistic content of the stimuli, and signers with higher language levels, both DESs and ISs, were assessed as more fluent.

### 5.4.2. Automatic Classification of Signers

To answer RQ2.3, we analyzed the results of the multi-label signer classifier (Table D.7 in Appendix D). The multi-label classifier obtained an F1 score of 0.312, meaning that it was able to correctly identify a signer one time in three, based only on comprehension and fluency values, and on the total number of glosses, outperforming the total accuracy obtained by human raters.

Figure 2 displays the confusion matrix of the signer classifier, over a 10-fold cross-validation on all data. The overlap in identification between DESs and ISs that we described in Section 5.2.2 persists, but in this case it was the ISs that were most frequently mistaken for DESs. The greatest confusion was between Signers 6 and 1 as well as Signers 2 and 5.

### 5.4.3. Justification Analysis

Whenever raters indicated having identified a signer, they were asked to elaborate on the factors that had led to identification. This information allows us to go deeper into RQ2.2. We qualitatively investigated the identifying factors that we had coded in the justifications (Section 4.3).



Groups	DES	IS	LS	Total
DER	<b>23/60 (38.33%)</b>	<b>21/60 (35.0%)</b>	<b>9/40 (22.5%)</b>	<b>53/160 (33.12%)</b>
IR	5/55 (9.09%)	18/59 (30.51%)	0/49 (0.0%)	23/163 (14.11%)
LR	25/195 (12.82%)	2/155 (1.29%)	14/185 (7.57%)	41/535 (7.66%)
<b>Total</b>	<b>53/310 (17.1%)</b>	<b>41/274 (14.96%)</b>	<b>23/274 (8.39%)</b>	<b>117/858 (13.64%)</b>

Table 3: Number of correct identifications (percentages in brackets) across language groups; without unknown familiarity. Values in bold indicate the highest accuracy scores for each signer group.

	Signers DE			4	Signers I		7	Signers L	
	1	2	3		5	6		8	9
Raters DE	7/20 (35.0%)	<b>9/20 (45.0%)</b>	7/20 (35.0%)	6/20 (30.0%)	11/20 (55.0%)	4/20 (20.0%)	<b>8/10 (80.0%)</b>	0/15 (0.0%)	1/15 (6.67%)
Raters I	1/18 (5.56%)	4/19 (21.05%)	0/18 (0.0%)	2/20 (10.0%)	<b>14/19 (73.68%)</b>	2/20 (10.0%)	0/19 (0.0%)	0/15 (0.0%)	0/15 (0.0%)
Raters L	9/65 (13.85%)	0/65 (0.0%)	16/65 (24.62%)	0/65 (0.0%)	2/65 (3.08%)	0/25 (0.0%)	13/65 (20.0%)	1/60 (1.67%)	0/60 (0.0%)

Table 4: Number of correct identifications (percentages in brackets) per rater group for each signer. Identification numbers in bold represent signers who were also raters. Values in bold highlight the signer within each signer group who received the highest identification rate.

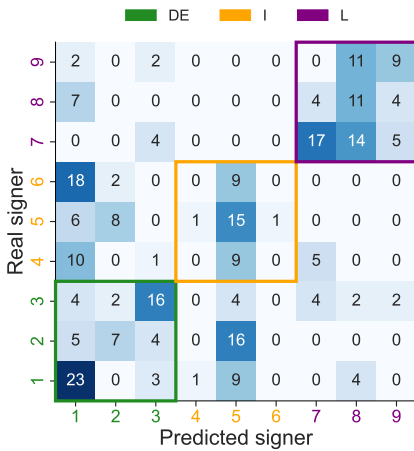


Figure 2: Confusion matrix for the signer classifier, evaluated using 10-fold cross-validation. The colored box indicates the language level of the signers.

Figure 3 shows the frequency distribution of the factors in each group of signers. In general, the factors focused on intrinsic characteristics of signers, such as the use of specific non-manual components or posture. Only a few raters indicated a non-descriptive factor, such as work, as an identifying feature.

For each group of signers, we characterized the main identifying features. The most important factors in identifying DESs were *signing style*, *posture*, *signing fluidity*, and non-manual components such as *head movements*. For instance, Rater 8’s observation of Signer 1 was as follows: “I can recognize them by the facial expression, positioning of the head, by the way they move the mouth, and by the fluidity of their signing.”

ISs were mostly assigned a *signing style* label, followed by the labels *grammatical aspects*, *mouth movement*, and *posture*. The *signing style* feature

may be attributed to the fact that the interpreters chosen as signers work for the national broadcaster and raters were familiar with seeing them on television. Regarding Signer 5, Rater 8 remarked, “They are recognizable by the look towards the monitor, by the signing speed, and by the movement of the body. This person uses many mouth actions. Also knowing how to meaningfully formulate the sentence content. Syntax is heavily influenced by German syntax. All this is typical of TV interpreters.”

For the LSs, *work* interactions were often mentioned as identifying reasons, indicating that raters who correctly identified LS were familiar with their signing style due to encounters in a work environment. The *work* code was used to label both the teacher-student and student-student relations that were indicated in the justifications. *Gesture* and *movements of the mouth* were cited as further identifying features. Rater 9 stated on Signer 7 that they were identifiable from “the way this person signs the word NAME and the excessive way they use the movements of the mouth.”

Finally, we explored the self-identification cases. Five out of the six raters who also appeared as signers successfully identified themselves and explicitly stated this in their justifications. Rater 15 briefly explained that they identified themselves based on their movements. These statements broadly demonstrate a certain degree of self-awareness regarding the raters’ own movement or movement in the action performed, a phenomenon previously observed (Loula et al., 2005; Bläsing and Sauzet, 2018).

## 6. Discussion

The rising concern for the privacy of sign language users, particularly in smaller deaf communities, prompted our study to inspect the assumption that pose estimates are anonymous representations

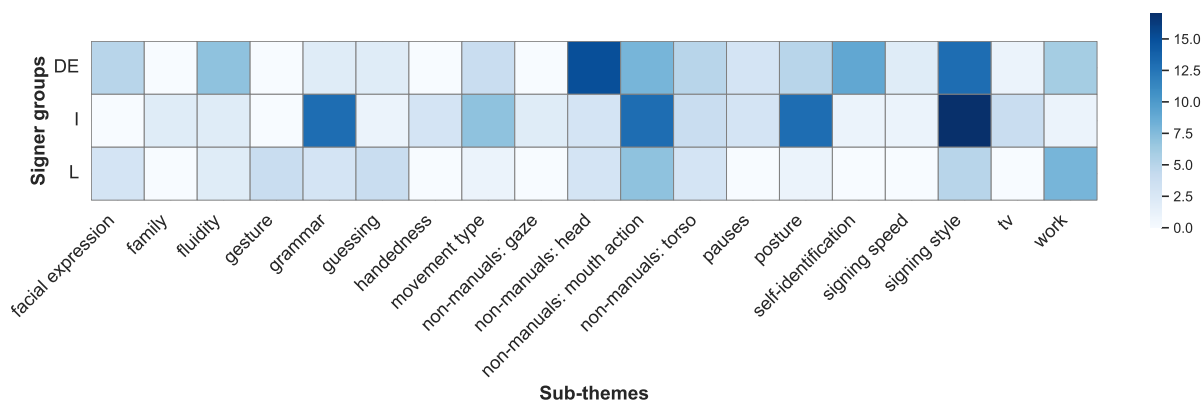


Figure 3: Matrix of the distribution of identifying factors across signer groups.

of sign language data. Contrary to this assumption, our findings reveal that participants were able to determine both the signer’s language level and identity with a certain degree of accuracy.

Automation of identification tasks, simulating potential applications in SLP, showed high F1 scores, indicating that non-anonymized DSGS pose sequences could be correctly identified at least one out of every three times. This result alone should raise concerns regarding the sharing and utilization of data without proper anonymization.

Our investigation also explored the role of subjective comprehension and fluency as predictors for identification tasks. The differences between the results obtained by the raters and the classifiers (e.g., Table 2 vs. Table D.6) prove that human raters leverage some additional features during the identification process that we did not collect with our survey, and thus could not be replicated by the classifiers.

Qualitative analysis of justifications highlighted factors like familiarity, movement, and signer-group-specific characteristics contributing to identification accuracy. Specifically, movement proved to be an identifying factor, aligning with existing studies in kinematics.

Considering the privacy concerns of sign language users, often hesitant to participate in research, our study emphasizes the need for anonymization methods, both at the visual appearance and individual motion levels. Striking a balance between data usefulness and privacy preservation is crucial as the field of SLP expands. While transforming sign language datasets into anonymized pose estimates presents a potential solution, its integration with novel systems and the acceptance of these strategies in sign language communities remain unexplored.

Acknowledging limitations such as the small participant pool and potential impacts of cultural and educational backgrounds, our findings stress the necessity of ongoing efforts to ensure the well-

being and protection of sign language users in the evolving landscape of sign language research.

## 7. Acknowledgments

This work was funded through the Swiss National Science Foundation (SNSF) Sinergia project “Scalable Multimodal Sign Language Technology for Sign Language Learning and Assessment II” (SMILE II) (grant agreement no. CRSII5\_193686) and the Swiss Innovation Agency (Innosuisse) flagship ICT (PFFS-21-47). In addition, the authors would like to thank Zifan Jiang for extracting the poses, and Arthur Capozzi for his precious suggestions on the analyses and valuable feedback.

## 8. Bibliographical References

- Jennifer S. Beal-Alvarez. 2016. [Longitudinal Receptive American Sign Language Skills Across a Diverse Deaf Student Body](#). *The Journal of Deaf Studies and Deaf Education*, 21(2):200–212.
- Félix Bigand, Elise Prigent, and Annelies Braffort. 2020. [Person identification based on sign language motion: Insights from human perception and computational modeling](#). In *Proceedings of the 7th International Conference on Movement and Computing*, MOCO '20, New York, NY, USA. Association for Computing Machinery.
- Félix Bigand. 2021. *Extracting human characteristics from motion using machine learning : the case of identity in Sign Language*. Ph.D. thesis, Université Paris-Saclay.
- Julian Bleicken, Thomas Hanke, Uta Salden, and Sven Wagner. 2016. [Using a language technology infrastructure for German in order to anonymize German Sign Language corpus data](#).

- In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3303–3306, Portorož, Slovenia. European Language Resources Association (ELRA).
- Bettina E. Bläsing and Odile Sauzet. 2018. [My action, my self: Recognition of self-created but visually unfamiliar dance-like actions from point-light displays](#). *Frontiers in Psychology*, 9.
- Penny Boyes Braem, Tobias Haug, and Patty Shores. 2012. Gebärdenspracharbeit in der Schweiz: Rückblick und Ausblick. *Das Zeichen*, 90:58–74.
- Danielle Bragg, Naomi Caselli, Julie A. Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E. Ladner. 2021. [The fate landscape of sign language ai datasets: An interdisciplinary perspective](#). *ACM Trans. Access. Comput.*, 14(2).
- Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. [Exploring collection of sign language datasets: Privacy, participation, and model performance](#). In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '20*, New York, NY, USA. Association for Computing Machinery.
- Hannah Bull, Michèle Gouiffès, and Annelies Brafort. 2020. [Automatic Segmentation of Sign Language into Subtitle-Units](#). In *Computer Vision – ECCV 2020 Workshops*, Lecture Notes in Computer Science, pages 186–198, Cham. Springer International Publishing.
- Necati Cihan Camgoz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. [Content4all open research sign language translation datasets](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5, Los Alamitos, CA, USA. IEEE Computer Society.
- Zhe Cao, Ginés Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. [Openpose: Real-time multi-person 2d pose estimation using part affinity fields](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(01):172–186.
- Zhe Cao, Ginés Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. [Openpose: Realtime multi-person 2d pose estimation using part affinity fields](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Council of Europe. 2009. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.
- Onno Crasborn. 2008. [Open access to sign language corpora](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 33–38, Marrakech, Morocco. European Language Resources Association (ELRA).
- Mirella De Sisto, Dimitar Shterionov, Irene Murtagh, Myriam Vermeerbergen, and Lorraine Leeson. 2021. [Defining meaningful units. Challenges in sign segmentation and segment-meaning mapping \(short paper\)](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 98–103, Virtual. Association for Machine Translation in the Americas.
- Sylvie Gibet. 2018. [Building French Sign Language motion capture corpora for signing avatars](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 53–58, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ivan Grishchenko and Valentin Bazarevsky. 2020. [MediaPipe Holistic](#). <https://blog.research.google/2020/12/mediapipe-holistic-simultaneous-face.html>.
- Matthew L Hall and Jess A Reidies. 2021. [Measuring Receptive ASL Skills in Novice Signers and Nonsigners](#). *The Journal of Deaf Studies and Deaf Education*, 26(4):501–510.
- Carlos Hinojosa, Juan Carlos Niebles, and Henry Arguello. 2021. [Learning privacy-preserving optics for human pose estimation](#). In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2553–2562.
- Amy Isard. 2020. [Approaches to the Anonymisation of Sign Language Corpora](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 95–100, Marseille, France. European Language Resources Association (ELRA).
- Amy Isard and Reiner Konrad. 2022. [MY DGS – ANNIS: ANNIS and the Public DGS Corpus](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*,

- pages 73–79, Marseille, France. European Language Resources Association (ELRA).
- Jindong Jiang, Wafa Skalli, Ali Siadat, and Laurent Gajny. 2022. [Effect of face blurring on human pose estimation: Ensuring subject privacy for medical and occupational health applications](#). *Sensors*, 22(23).
- Trevor Johnston. 2004. [The assessment and achievement of proficiency in a native sign language within a sign bilingual program: the pilot auslan receptive skills test](#). *Deafness & Education International*, 6(2):57–81.
- Annelies Maria Jozef Kusters, Dai O'Brien, and Maartje De Meulder. 2017. *Innovations in Deaf Studies: Critically Mapping the Field*, pages 1–53. Oxford University Press, United Kingdom.
- Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. [American sign language video anonymization to support online participation of deaf and hard of hearing users](#). In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '21, New York, NY, USA. Association for Computing Machinery.
- Fani Loula, Sapna Prasad, Kent Harber, and Maggie Shiffrar. 2005. [Recognizing people from their movement](#). *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):210–220.
- Kelly Mack, Danielle Bragg, Meredith Ringel Morris, Maarten W. Bos, Isabelle Albi, and Andrés Monroy-Hernández. 2020. [Social app accessibility for deaf signers](#). *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).
- Liliane Momeni, Hannah Bull, K. R. Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. [Automatic Dense Annotation of Large-Vocabulary Sign Language Videos](#). In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, volume 13695, pages 671–690. Springer Nature Switzerland, Cham. Series Title: Lecture Notes in Computer Science.
- Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgoz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023. [Findings of the second WMT shared task on sign language translation \(WMT-SLT23\)](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.
- Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgoz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. [Findings of the first WMT shared task on sign language translation \(WMT-SLT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Jemina Napier and Lorraine Leeson. 2016. [Sign Language in Action](#). In *Sign Language in Action*, pages 50–84. Palgrave Macmillan UK, London.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#).
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. [Anonymsign: Novel human appearance synthesis for sign language video anonymization](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, page 1–8. IEEE Press.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. [Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production](#). ArXiv:2203.15354 [cs].
- Marc Schuler and Thomas Hanke. 2020. [OpenPose in the Public DGS Corpus](#). Publisher: Universität Hamburg Version Number: 2.
- Mai Skjott Linneberg and Steffen Korsgaard. 2019. [Coding qualitative data: a synthesis guiding the novice](#). *Qualitative Research Journal*, 19(3):259–270. Publisher: Emerald Publishing Limited.
- Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. [Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks](#). *International Journal of Computer Vision*, 128(4):891–908.
- Christina O. Tze, Panagiotis P. Filntisis, Anastasios Roussos, and Petros Maragos. 2022. [Cartoonized Anonymization of Sign Language Videos](#). In *2022 IEEE 14th Image, Video, and*



*Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, Nafplio, Greece. IEEE.

Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. [Read and Attend: Temporal Localisation in Sign Language Videos](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16852–16861, Nashville, TN, USA. IEEE.

Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitri Metaxas. 2022. [Sign language video anonymization](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 202–211, Marseille, France. European Language Resources Association.

Zhaoyang Xia, Carol Neidle, and Dimitris N. Metaxas. 2023. [DiffSLVA: Harnessing Diffusion Models for Sign Language Video Anonymization](#). ArXiv:2311.16060 [cs].

## A. Example Anonymization Methods

Figure A.1 shows six examples of techniques applied in research to (pseudo-)anonymize sign language data (Section 2).

## B. Annotation Codebook

Theme	Sub-themes
Non-manuals Signing	mouth, gaze, eyebrows, head, torso signing style, gesture, handedness, grammar, posture
Self-identification	self-identification
Movement	movement type
Fluency	signing fluidity, pauses, signing speed
Appearance	body, facial expression
Other	work, TV, family, guessing

Table B.5: Codebook containing themes and sub-themes identified in the justifications. Note that sign language movement was coded as *movement*, while upper body movement was annotated using the code *non-manuals: torso*.

## C. Familiarity

Figure C.2 shows the results of the follow-up survey, in which each rater was required to indicate their familiarity with each signer using a “yes” or “no” response (Section 3.3). The three DESs were known by all raters, while there is a degree of variability regarding the reported familiarity for the two other groups of signers, especially for the LSs.

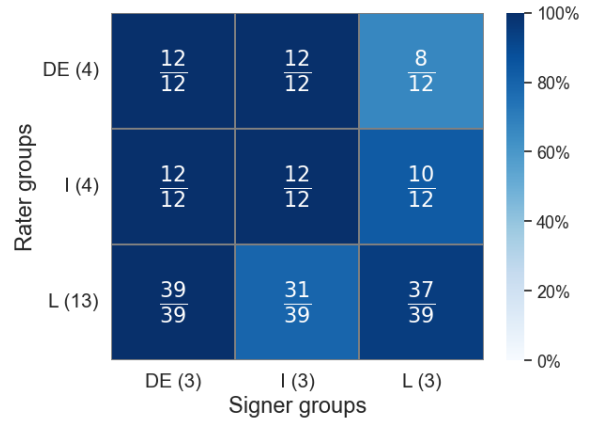


Figure C.2: Plot comparing the familiarity of signers across raters. Values in brackets indicate the number of persons in the group. Values within the cells denote the proportion of familiarity between the raters and the signers, while the color gradient indicates the corresponding percentage.

## D. Classifier Results

Table D.6 reports the results for the “language classifier” described in Section 5.2.2. Table D.7 presents the results for the “signer classifier”, described in Section 5.4.2.

	Precision	Recall	F1	Accuracy
DE	0.606	0.588	0.559	0.594
I	0.778	0.811	0.776	0.784
L	0.847	0.866	0.852	0.865
Dummy DE+I+L	0.112	0.333	0.168	0.336
DE+I+L	0.645	0.657	0.638	0.657

Table D.6: Average scores for the binary classifier, dummy multi-label classifier, and multi-label language classifier, evaluated with a 10-fold cross-validation. DE+I+L is the final classifier.

	Precision	Recall	F1	Accuracy
Dummy	0.012	0.111	0.022	0.108
Signer	0.342	0.336	0.312	0.336

Table D.7: Average scores for the dummy multi-label signer classifier and multi-label signer classifier, evaluated with a 10-fold cross-validation.

## E. Plots RQ1

Figures E.3, E.4, and E.5 are visualizations discussed in Section 5.2, concerning RQ1 on identifying the language level of signers.

## F. Plots RQ2

Figures F.6 and F.7 are visualizations described in Section 5.4 regarding RQ2 on person identification.

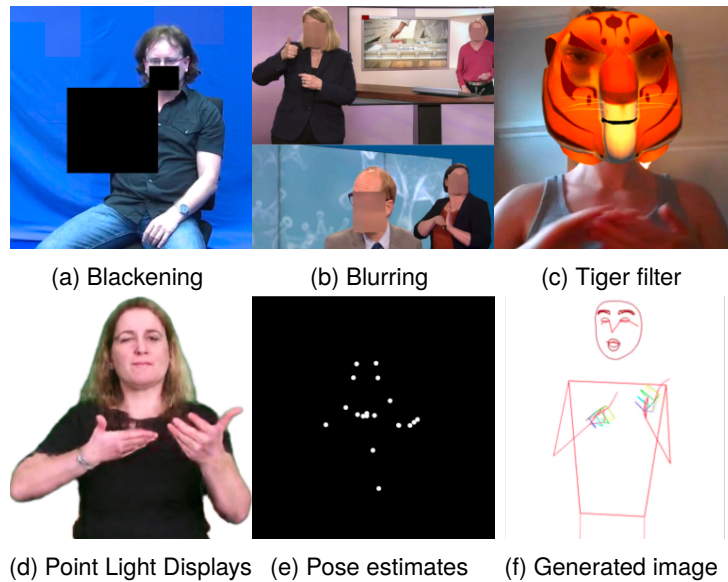


Figure A.1: Examples of methods used for anonymizing sign language data. Picture (a) from (Isard, 2020); picture (b) from (Camgoz et al., 2021); picture (c) from (Bragg et al., 2020); picture (d) from (Saunders et al., 2021); picture (e) from (Bigand et al., 2020); picture (f) from our study.

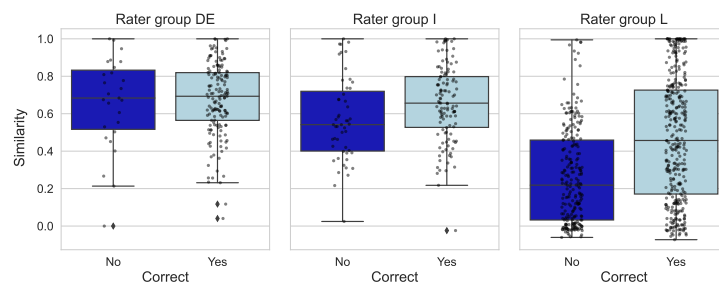


Figure E.3: Distribution of similarity scores for correct and incorrect identifications of the signers' language levels, across rater and signer groups.

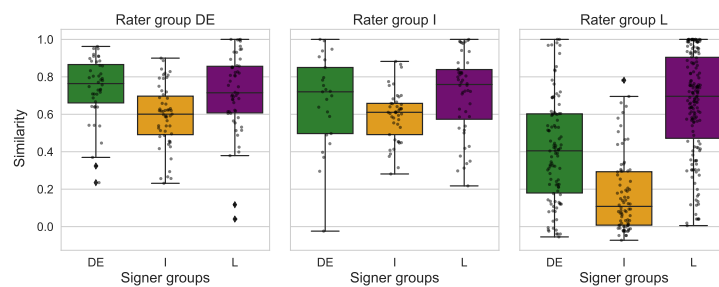


Figure E.4: Distribution of similarity scores for correctly identified language levels across signer groups. Each subplot corresponds to a different rater group and illustrates the distribution of similarity values (on the y-axis) obtained by rater groups in transcribing the content of the utterances from videos where they correctly identified the language levels of the signers.

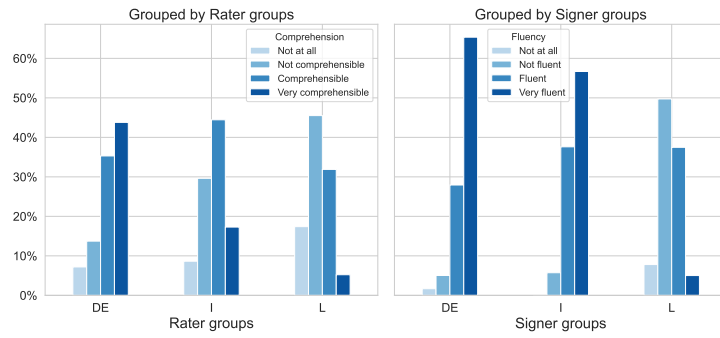


Figure E.5: Left: Bar plot showing the distribution of comprehension levels among rater groups. The y-axis represents percentages and the x-axis displays the four comprehension values across the rater groups. Right: Bar plot showing the distribution of fluency ratings among three signer groups. The y-axis represents percentages, and the x-axis displays the three signer groups and the four assigned fluency ratings, ranging from *Not at all fluent* to *Very fluent*.

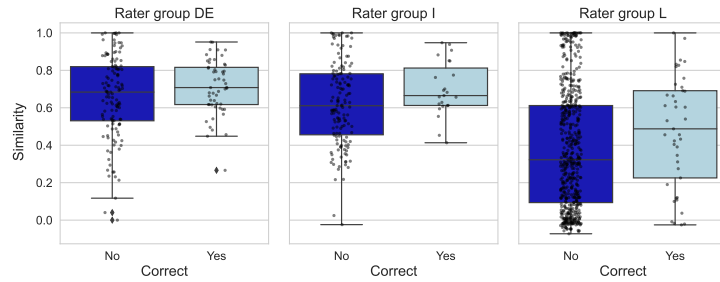


Figure F.6: Distribution of similarity scores for correct and incorrect signer identifications, across rater and signer groups.

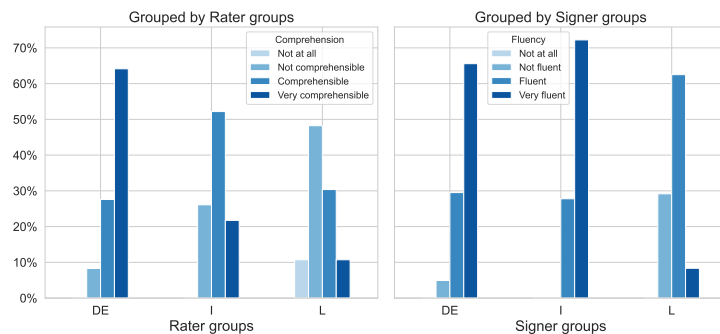


Figure F.7: Right: Bar plot showing the distribution of the comprehension ratings assigned by the raters to the stimuli whose signers were correctly identified. Left: Bar plot showing the distribution of fluency ratings among three signer groups.

# Data Integration, Annotation, and Transcription Methods for Sign Language Dialogue with Latency in Videoconferencing

Mayumi Bono<sup>1&2</sup>, Tomohiro Okada<sup>1</sup>, Victor Skobov<sup>2</sup>, and Robert Adam<sup>3</sup>

<sup>1</sup> National Institute of Informatics, <sup>2</sup> SOKENDAI (The Graduate University of Advanced Studies),  
<sup>3</sup> Heriot-Watt University

<sup>1</sup> 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN, <sup>2</sup> Shonan Village, Hayama, Kanagawa  
240-0193 JAPAN, <sup>3</sup> Edinburgh, Scotland EH14 4AS  
{bono, tokada-deaf, vskobov}@nii.ac.jp, R.Adam@hw.ac.uk

## Abstract

This article aims to explain how latency is captured in sign language dialogue via videoconferencing and how recorded data are integrated and annotated using an annotation tool (ELAN). First, we present two examples of the analysis to clarify basic theoretical issues that affect turn-taking via videoconferencing systems focusing on the sequence structure of ‘greetings’ and ‘encounters.’ Videoconferencing dialogues often begin with the participants greeting each other, which may be delayed because of the nature of online communication or the technical specifications of each individual’s device. Next, to discuss sequential issues with videoconferencing dialogue, we introduce how the fundamental adjacency pair, such as question (first pair part: FPP) and answer (second pair part: SPP), appears to each participant on their computers with latency. This research shows that recording videoconferencing dialogues with latency is useful for next-generation data collection in vision-sensitive sign languages, as well as audio-centred spoken languages with gestures.

**Keywords:** latency, videoconferencing, sign language dialogue

## 1. Introduction

This article aims to explain how latency is captured in sign language dialogue via videoconferencing and how recorded data are integrated and annotated using an annotation tool (ELAN). Since the start of the coronavirus disease 2019 (COVID-19) pandemic, online conferencing has become a part of daily life for many people. This lifestyle change applies to hearing people and Deaf people. How have Deaf individuals, who essentially communicate in three-dimensional space, experienced this shift? To address this question, the present study recorded online conversations between Deaf people using the videoconferencing tool Zoom.

Before the coronavirus disease 2019 (COVID-19) pandemic, Deaf people would meet in so-called Deaf spaces, where they could communicate using sign language—thus, they formed their own society (Kusters, 2015). The pandemic forced Deaf people to meet online, and the Deaf community, which values face-to-face communication, was inspired to extend Deaf space into two-dimensional spaces such as videoconferencing. The long COVID-19 pandemic facilitated human familiarity with and adoption of videoconferencing systems in daily life, resulting in a stable world where Deaf people worldwide can communicate across spatial and distance barriers. Deaf people have been using videoconferencing before COVID-19, and it has been reported that they have unique linguistic

and ethnographic ways of integrating such new technologies into their lives (Keating and Mirus, 2003). Before Corona, the Deaf who participated in online communication were a small group of people with strong computer skills, and their use was not stable and continuous. The increase in use and adaptation of online communication in the wake of the coronavirus disaster raises long-term observation needed theoretical questions in Communication Studies regarding the effects on how Deaf people, who have essentially communicated in three-dimensional space, communicate with others in two-dimensional digital space via sign language<sup>1</sup>.

In terms of linguistic resources for natural language processing research, videoconference recordings of dialogues could be useful for next-generation data collection. Data recording using videoconferencing systems, which do not require participants to meet in person, will prevent the spread of unknown viruses in the future and allow data recording by people from different regions. For example, the geographic distance between the UK and Japan meant that contact between their respective sign languages was impossible in face-to-face situations. However, now that online communication is commonplace, Deaf people in the UK and Japan can meet more easily and frequently than before.

Here, we report the preliminary results of part of the 3-year international joint project ‘Understanding cross-signing phenomena in video conferencing situations during and post-

<sup>1</sup> There are already projects documenting the experiences of Deaf communities in the time of COVID-19 for American Sign Language.

<https://doi.org/10.6084/m9.figshare.22340830.v1>



COVID-19 in rural areas'<sup>2</sup> between the United Kingdom (UK) and Japan, which began in 2022. The goal of this project is to observe online cross-signing phenomena among non-shared language situations (Bono and Adam, 2023); it consists of two phases. In the first phase (2022/23), data collection was conducted in the respective countries (UK and Japan) using videoconferencing systems. During the second phase (2023/24), Deaf people in Japan and the UK, who do not have a shared sign language, will meet and interact with each other through a videoconferencing system.

In this article, we describe data integration, annotation, and transcription methods for video clips with videoconferencing-specific latency that were designed by the Japanese team during the first phase. First, we present two examples of the analysis to clarify basic theoretical issues that affect turn-taking via videoconferencing systems focusing on the sequence structure of 'greetings' and 'encounters.' Videoconferencing dialogues often begin with the participants greeting each other, which may be delayed due to the nature of online communication or the technical specifications of each individual's device. To discuss theoretical issues with videoconferencing dialogue, we introduce how the fundamental repair sequence, such as question and answer, appears to each participant on their local computers with latency. This research helps to show that recording videoconferencing dialogues with latency is useful as next-generation data collection for vision-sensitive sign languages, as well as audio-centred spoken languages with gestures.

Section 2 describes the methods used to process the delays; section 3 gives an overview of the data collection; and section 4 demonstrates the actual qualitative analysis of the data. This paper is the first report to show how latency is essential for qualitative analysis research on online sign language dialogues.

## 2. Latency in Videoconferencing

From a technical perspective, many videoconferencing systems seek lower latency to more closely resemble in-person conversations. However, depending on internet speeds and computer specifications, latency may be high in an individual's home. Many sociological and conversation analytical studies of video-mediated interactions have focused on the lack of shared space in conversations that occur via videoconferencing systems (Heath and Luff, 1993). Even in spoken conversation, if the space is not shared, it becomes difficult to use gestures such as eye contact and pointing, which can typically be used without difficulty during face-to-

face interactions. In the aftermath of the COVID-19 pandemic, Seuren et al. (2021) observed a remote medical interview conducted using Skype<sup>3</sup>, which had been the predominant videoconferencing platform before COVID-19—rather than Zoom<sup>4</sup>—using the Conversation Analysis (CA) method. They concluded that conversation participants communicating via videoconferencing platforms behave as though they inhabit a shared reality.

We believe that two issues must be considered here. The first issue is the importance of latency in interactions such as medical counselling, where the goal is 'solving' or 'curing' a problem. During social interactions, in which the explicit goal is achievement of the objective regardless of latency or transmission problems, these problems may be tolerated if the goal is achieved. The second issue arises in situations where Deaf people use videoconferencing systems. When hearing people use videoconferencing systems, they have the option to cease using the video component if latency or video outages occur; however, Deaf individuals do not have that option. Additionally, Zoom has a function that—if the audio transmission ceases for a certain period of time—allows users to increase the audio speed and transmit all speech that can be understood and heard. Conversely, Zoom does not have a function to reduce the video frame rate and transmit language-understandable and readable video in a single transmission. Thus, when latency or video outages occur, the Deaf person must be able to clearly resolve these troubles so that they can follow the conversation.

The 'greeting' and 'encounter' situations in online communication are the first places where latency due to the recipient's internet environment and personal computer specifications can be identified. If latency in the recipient's video transmission is recognised, it will be necessary for the speaker to consider such latency. When discussing delays in online communication, it is important to discuss this system-induced trouble, which can be termed 'basal latency'. Basal latency results in different ways of viewing sequence organisation between oneself and others in a videoconferencing dialogue. In this paper, we focus on basic adjacency pairs such as question (first pair part: FPP) and answer (second pair part: SPP) and raise theoretical issues regarding sequence organisation in CA (Schegloff, 2007).

## 3. Data Collection

The details of data collection during the first phase have been published elsewhere (Bono and Adam, 2023). This section introduces the method of data collection, focusing on latency and

<sup>2</sup> <https://www.ukri.org/news/uk-japanese-collaboration-to-address-covid-19-challenges/>

<sup>3</sup> <https://www.skype.com/en/>

<sup>4</sup> <https://zoom.us/>

components of the analysis detailed in Sections 5 and 6. Participants were selected from three geographically distant regions in Japan: Hokkaido, Shikoku, and Okinawa. Three participants were selected from each of the abovementioned regions, and then divided into groups A, B, and C for each region (see Table 1). Dialogue pairs were composed of one participant from each group and the other participant from one of the remaining two groups—for example, the ‘Hokkaido (HK) and Shikoku (SK) pair’, the ‘SK and Okinawa (ON) pair’, and the ‘HK and ON pair’ in Group A.

Region	Group A /ID	Group B /ID	Group C /ID
Hokkaido	HK-A	HK-B	HK-C
Shikoku	SK-A	SK-B	SK-C
Okinawa	ON-A	ON-B	ON-C

Table 1: Regions, groups, and identifications (IDs)

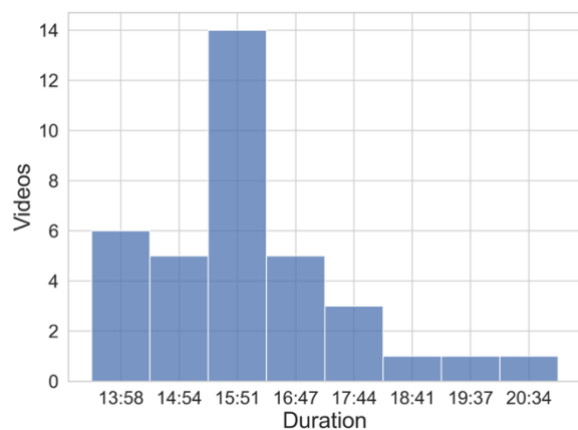


Figure 1: Video Duration Distribution

The online dialogue was recorded locally on the participants’ computers using the recording function in Zoom at three sites: the locations of both participants and the monitoring staff (Zoom host). Using the ‘hide non-video participants’ feature in Zoom, the monitoring staff faded from the Zoom view of the participants as the conversation/experiment began. However, the monitoring staff actually participated in the Zoom call to gauge and monitor the participants’ dialogues. The reason for recording at each site was to avoid missing any discussion of latency issues during online communication that might have affected the turn-taking process (Seuren et al., 2021). By recording at three sites, it was possible to process and analyse the timings of various communication phenomena; this allowed the researchers to determine how each participant saw their recipient’s image and

identify any differences in the way they might subsequently view each other.

After the monitoring staff member turned off their camera and appeared to have left the session, the participants commenced their online dialogue. At the appropriate time, as the conversation was ending (e.g., as indicated by topic shifts; approximately 15 minutes), the monitoring staff member would turn on their camera to terminate the ongoing dialogue. Figure 1 illustrates the distribution of the video durations, showing that most dialogues concluded within approximately 15 minutes but sometimes continued for up to 20 minutes.

## 4. Latency in Analysis

Latency has a noticeable impact on the conversation process: a certain degree of latency can make the conversation impossible. Thus, this study tracked latency during the data collection process.

### 4.1 Capturing Latency in Zoom

Latency has a noticeable impact on participant satisfaction with the conversation process. If the delay reaches 400 ms, the conversation will become unacceptable for participants (ITU-T, 1996). Garg et al. (2022) reported that participants were able to adapt to higher latency, but they exhibited increased fatigue and frustration associated with higher cognitive load during visual tasks. In the context of data collected from sign language dialogues held via videoconferencing, latency tracking and reporting are essential for future conversation analyses.

The built-in tools for latency tracking and reporting in Zoom have an ambiguous description<sup>5</sup> and unclear export capabilities; a requirement for participants to use these tools would add unwanted complexity to the recording process. For post-collection latency measurement, we chose a three-way setup—two participants and a monitor—as shown in Figure 2.

Using this setup, the delay between the two participants could be fully observed only by a monitoring party. The observation was also shifted along the absolute timeline because the observer had its delay. Nonetheless, this observation added context to each participant’s recordings, allowing us to synchronise them within the absolute timeline.

Zoom has a function that allows conversations to be recorded and stored in the cloud or in the local memory. The difference between the two options is crucial: if a participant records to the cloud, a

<sup>5</sup>Available at: <https://support.zoom.us/hc/en-us/articles/202920719-Accessing-meeting-and-phone-statistics>

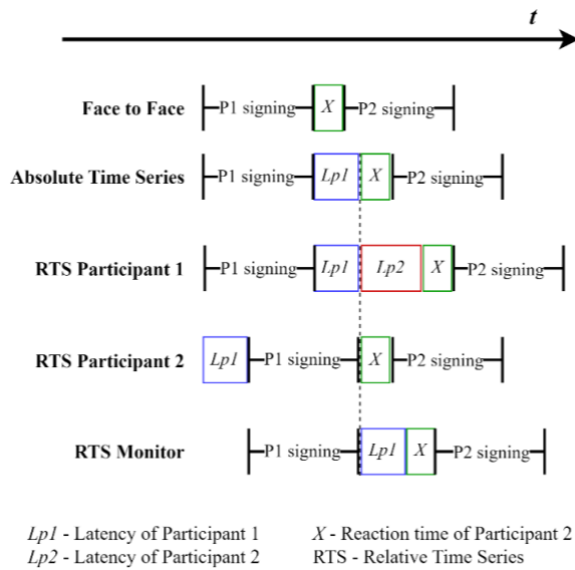


Figure 2: Three-way latency conversation schema, adapted with permission from Hosoma and Muraoka (2022)

delay will be added to his camera view, and the video quality will be reduced. Recordings stored in local memory have superior quality and no delay; thus, this storage is a critical requirement for post-collection latency computation.

The synchronisation is performed by calculating the time shift between the participants' and monitor's records. The participants' recordings are trimmed accordingly, after which they begin simultaneously in the absolute timeline and are effectively synchronised with the monitor's record. They may then be used to measure latency between participants.

The latency and synchronisation time shifts were calculated using cross-correlation within SciPy<sup>6</sup>. For this purpose, we reduced each video to a one-dimensional signal by calculating the Euclidean distance between each frame and an average frame of the entire video.

Participants' recordings were compared with the received version in the other recordings. Each corresponding piece of the frame with the participant's view was cropped to the view size prior to calculation. For synchronisation with the monitor's record, we collected a small portion at the same video position (250 frames). A sliding

<sup>6</sup> Available at: [https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.correlation\\_lags.html#scipy.signal.correlation\\_lags](https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.correlation_lags.html#scipy.signal.correlation_lags)

<sup>7</sup> This ELAN annotation is a preliminary step before the ELAN integration adjustment method is applied based on absolute time, as described in Section 4.2. In this context, M-view means monitoring view, HK-view means Hokkaido view, and ON-view means Okinawa

window of 120 frames was used to determine participant latency at each frame.

## 4.2 ELAN Integration

ELAN Software, which is used to annotate the sign language corpus, has a built-in function that allows time series to be displayed along the video timeline. We utilised this functionality to display the calculated latency in the recordings, as illustrated in Figure 3. The output facilitates comprehension of the delay and reaction time.

Delayed annotations may be created by adding latency to the start and end times in the annotation within the absolute timeline. This addition may be done automatically using the Python *pympl-ing* module.

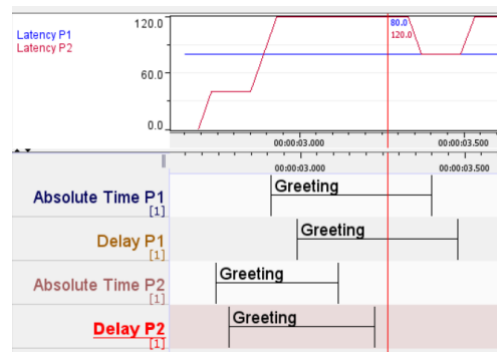


Figure 3: Latency display in ELAN

## 5. Analysis of "Greetings" and "Encounters" in Videoconferencing

### 5.1 Analysis 1: Sequential "Hi" or floating "Hi" (Hokkaido–Okinawa)

Analysis 1 focuses on dialogue of greeting scenes between Hokkaido and Okinawa in Group B (hereafter HK for the Hokkaido participant and ON for the Okinawa participant). Observing the results annotated with ELAN in Figure 4,<sup>7</sup> a sequential relationship can be identified in the monitoring view (recorded in Tokyo) and the Hokkaido view, where HK says 'Hi' first; ON then responds, 'Nice to meet you'.<sup>8</sup> Conversely, in the Okinawa view, it appears that ON said the words 'Nice to meet you' first, whereas HK said 'Hi' almost simultaneously (with a delay of

view on the ELAN tiers' names. Because the absolute time has not been adjusted, analysis between the different participant's views is impossible. Therefore, we compare the results between the same participant's views.

<sup>8</sup> Schegloff (2007) does not apply the concept of adjacency pairs to greeting sequences, so we follow this here and describe them as a 'sequential relationship' rather than adjacency pairs. We describe the concept of adjacency pairs in Section 6 more detail.

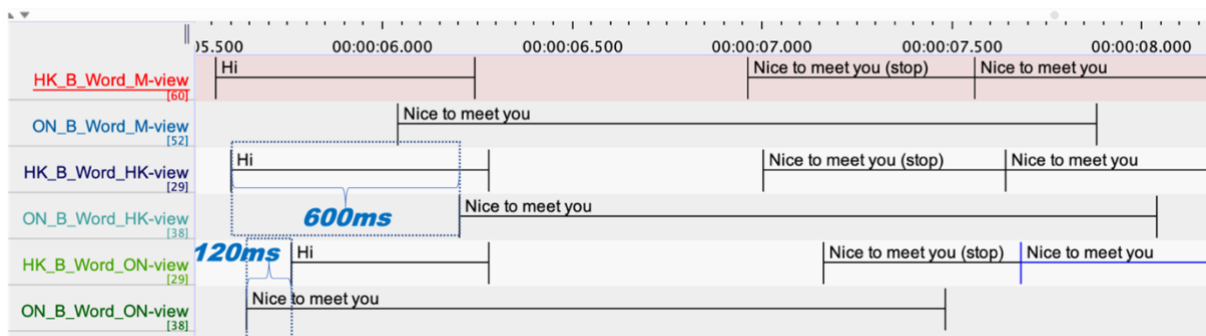


Figure 4: Analysis 1: Sequential “Hi” or floating “Hi” (Hokkaido–Okinawa)

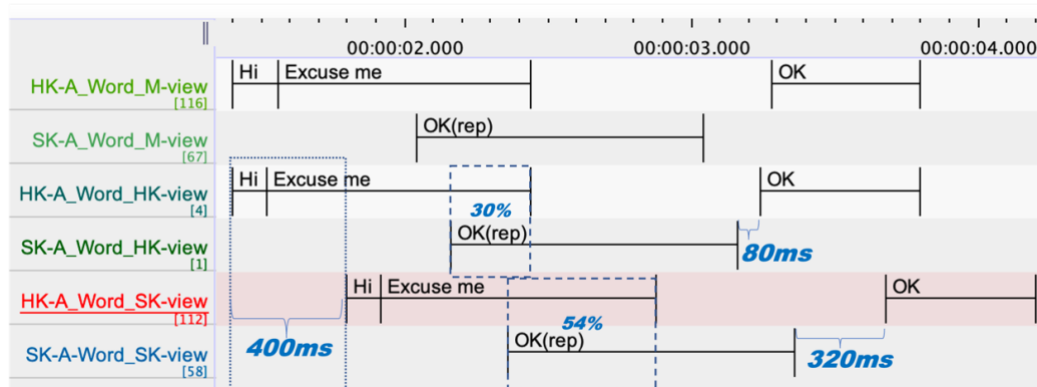


Figure 5: Analysis 2: Showing a positive attitude (Hokkaido–Shikoku)

approximately 120 ms from the ‘Nice to meet you’ by ON).

Next, in the transcript based on the CA notation, we described these differences (Excerpts 1 and 1’).<sup>9</sup> Theoretically, in CA, the monitoring and Hokkaido views indicate that the ‘Hi’ uttered by HK is in a sequential relationship with the ‘Nice to meet you’ next uttered by ON (see lines 01 and 02 in Excerpt 1). Conversely, in the Okinawa view, HK’s ‘Hi’ completely overlaps with ON’s ‘Nice to meet you’. In this scenario, it appears that ON’s ‘Nice to meet you’ is uttered first; HK then responds, ‘Nice to meet you’ (see lines 01 and 03 in Excerpt 1’). Accordingly, HK’s ‘Hi’ is considered to be floating in the sequence structure. Subsequently, it appears that HK says ‘Nice...Nice to meet you’ in line 03 following and imitating ON’s greeting in both Excerpts 1 and 1’. Thus, in a dialogue occurring via videoconferencing, HK and ON may hold completely opposite perceptions of who issued the first greeting.

As a part of the ELAN annotation in Figure 4, HK should feel that ON is responding 600 ms after the onset of his ‘Hi’ utterance. However, ON would have felt as though he had initiated his salutatory utterance 120 ms earlier than HK’s ‘Hi’. Simply adding these together, ON’s salutatory utterance is conveyed to HK with a delay of 720

ms. How does a delay of > 0.7 seconds (sec) affect the interaction? Analysis 2 continues the observation by examining another case.

#### Excerpt 1 (Monitoring and Hokkaido view)

01 HK: Hi  
02 ON: [Nice to [meet you  
03 HK: [Nice...Nice to meet you

#### Excerpt 1’ (Okinawa view)

01 ON: Ni[ce] to meet [you  
02 HK: [Hi  
03 HK: [Nice...Nice to meet you

## 5.2 Analysis 2: Showing a positive attitude

The data examined in Analysis 2 are derived from the beginning of the third dialogue experiment (Figure 5). It is an encounter, rather than a greeting, and HK initially apologises for his own connectivity problems. Similar to the data in Analysis 1, there is minimal latency between the monitoring view (recorded in Tokyo) and the Hokkaido view, but the recipient’s video transmission exhibits latency in the Shikoku view.

Simple observation of the beginning of ‘Hi’ uttered by HK in the Hokkaido view and Shikoku view indicates a basal latency of 400 ms between them. Further analysis reveals that SK’s ‘No worries (OK (rep)<sup>10</sup>)’ overlaps with the final 30% of HK’s

<sup>9</sup> The transcript of Excerpt 1 does not use the word glosses of the signs separated by slashes because this analysis does not aim to show temporal relations; it uses the Japanese translation.

<sup>10</sup> The signal of (rep) added after the word gloss means that the sign expression is repeated. Thus, [OK] is repeated several times here.

'Excuse me' utterances ('Excuse me': duration 920 ms, overlap time: 280 ms) in the Hokkaido view. In the Shikoku view, this percentage increases to 54% ('Excuse me': duration 960 ms, overlap time: 520 ms). SK's action in the Shikoku view, which overlaps by more than 50% with HK's utterance and responds to it, may be assumed to indicate a positive attitude towards the recipient. Accordingly, SK is repeatedly and quickly expressing 'No worries' to HK.

After SK's reply with repeated OK, HK closes the sequence by saying 'Alright' (sequence-closing 3rd). However, there is another difference between the Hokkaido and Shikoku views: in the Hokkaido view, HK closes SK's 'No worries' with 'Alright' without a pause (after a short gap of 80 ms). Conversely, in the Shikoku view, the transmission of HK's 'Alright' is delayed, and the sequence appears to terminate after a lengthy pause of 320 ms. Although this difference is minor, subtracting the actual gap of 80 ms from 320 ms results in a latency of 240 ms, indicating that HK's response, 'Alright' (sequence-closing 3rd), was not transmitted at the appropriate time. Thus, the influence of basal latency is present in these interactions. In the Shikoku-view, because of latency caused by the system, HK's reaction in line 04 has a weak relationship with the previous sequence, which is also floated from the fundamental sequence organisation.

#### Excerpt 2 (Hokkaido view)

01 HK: Hi/Excuse-[me/ (*Hi, Excuse me*)  
 02 SK: [OK (rep) (*No worries*)  
 03 (*gap: 80 ms*)  
 04 HK: OK (*Alright*)

#### Excerpt 2' (Shikoku view)

01 HK: Hi/Excuse-[me/ (*Hi, Excuse me*)  
 02 SK: [OK (rep) (*No worries*)  
 03. (*long pause: 320 ms*)  
 04 HK: OK (*Alright*)

In Excerpts 2 and 2' formed as a CA transcript, Excerpt 2 in the Hokkaido view sequentially appears better than Excerpt 2' in the Shikoku view; SK's response in line 02 terminally overlaps HK's apologies in line 01. Then, after an 80 ms gap, HK expresses 'Alright' (sequence-closing 3rd). In Excerpt 2', however, SK gives responses in line 02 with a positive attitude; there is no rapid sequential feedback from HK. In summary, this encounter is smooth for HK, whereas it is slightly awkward for SK.

## 6. Analysis of Sequence Organisation with Latency

As mentioned in footnote 7, Schegloff (2007) does not apply the concept of adjacency pairs to a sequence of greetings. Therefore, we should not analyse greetings or encountering; we should focus on the contents of the conversation

sequence after greetings to understand what occurs in an online dialogue with latency from the perspective of sequence organisation.

In Analysis 3, we focus on differences in the appearance of a simple question-answer adjacency pair between the two views. Analysis 4 shows how the theoretical issues raised in Analysis 3 may be treated in terms of the repair sequence (Kitzinger, 2013; Schegloff et al., 1977).

Recently, several researchers, mainly the language and cognition research group at the Max Planck Institute, have applied comparative and quantitative analysis to repair sequences, especially other-initiated repair (OIR), in several languages as a universal and fundamental system of human communication that transcend differences across cultures and communication modality, in spoken, signed, and tactile conversations (Bono et al., 2023; Byun et al., 2018; Dingemane and Enfield, 2015; Dingemane, Kendrick and Enfield, 2016; Dingemane, Torreira and Enfield, 2013; Floyd et al., 2016; Haakana et al., 2021; Hayashi et al., 2013; Kendrick, 2015; Manrique and Enfield, 2015; Manrique, 2016). This article focuses on more fundamental issues on CA such as adjacency pairs in Analysis 3, and self-initiated self-repair sequence not OIR in Analysis 4.

### 6.1 Analysis 3: Question-answer adjacency pairs

The data in Figures 6 and 7 were obtained from the first session, 26 s after the beginning. SK asks ON, LIVE/PLACE/WHERE, 'where do you live?' with questioning facial expressions. After the question, she maintains her hand shape and holds it in signing space, which is annotated as 'post-stroke-hold', while looking at the recipient. The concept of post-stroke-hold arises from Gesture Studies (McNeill, 1996; Kita et al., 1998; Kendon, 2004). In spoken conversation, post-stroke-hold functions to hold a topic in discourse, whereas it has several grammatical functions in sign language. Here, SK holds the conversational floor and connects her sequence-closing third, OKINAWA 'Okinawa (I see)', to line 03 in Figures 6 and 7. Sequence-closing thirds (SCTs) are placed in the third position of question-answer adjacency pairs by the person who asks a



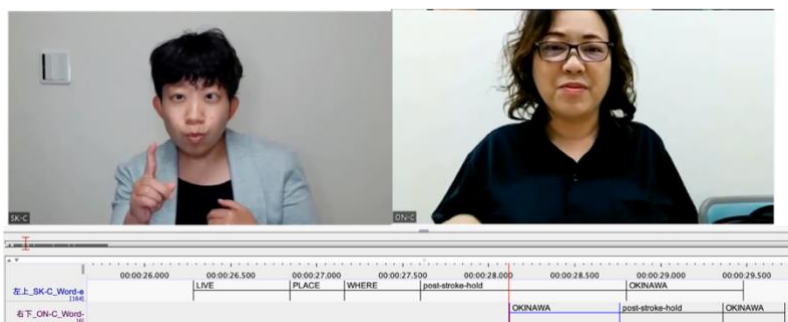


Figure 6: Q–A adjacency pair (Shikoku view)

- 01 SK-C: Where do you live?  
(0.5 s gap)
- 02 ON-C: (I live in) Okinawa
- 03 SK-C: Okinawa (I see)
- 04 ON-C: (I live in) Okinawa

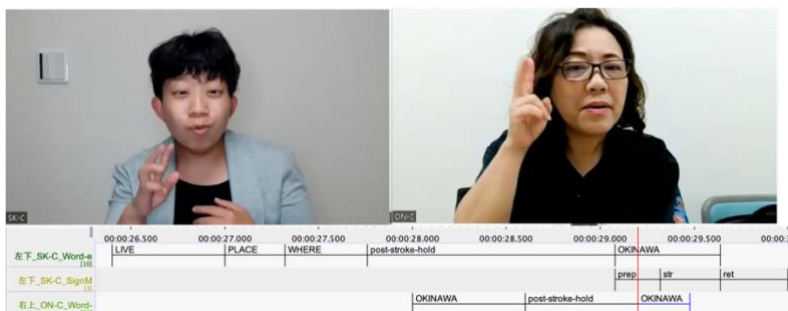


Figure 7: Q–A adjacency pair (Okinawa view)

- 01 SK-C: Where do you live?  
(0.2 s gap)
- 02 ON-C: (I live in) Okinawa  
(0.4 s gap)
- 03 SK-C: O[kinawa (I see)
- 04 ON-C: [Okinawa

question to evaluate the answer provided by the interlocutor and close the current adjacency pair.

During SK's post-stroke-hold, ON answers, OKINAWA, '(I live in) Okinawa'. There is a difference in the gap before answering between the views of Shikoku and Okinawa. In the Shikoku view, the gap is 0.5 s, whereas it is 0.2 s in the Okinawa view. We do not consider this a large difference in an adjacency pair.

The theoretical issue is the explanation for repetition of ON's answer in line 04. In the Okinawa view (Figure 7), the explanation is visible in SK's SCT, 'Okinawa (I see)', which arrives slightly later. There is a 0.4-s gap between

lines 02 and 03. Consequently, ON repeats the answer in line 04. We add more detailed sign movement annotations, prep (preparation), str (stroke), and ret (retraction) to SK's SCT (Kikuchi and Bono, 2013). From the detailed annotations, we observe that when ON begins the repetition, SK continues to prepare for OKINAWA as the SCT. In this context, we consider SK's reaction to ON's answer to be slightly delayed; subsequently, ON repeats her answer again in the Okinawa view. This is an example of self-initiated self-repair by ON (Schegloff et al., 1977; Kitzinger, 2013). ON notices her answer is not conveyed to the recipient, then tries her answer again.

In contrast, in the Shikoku view, SK's reaction is less delayed. SK begins the SCT, 'Okinawa (I see)', immediately after ON's answer. There is no gap here. This is the shortest time to close the sequence. Our question here is how ON's repetition in line 04 appears to SK.

First, some sign language linguists insist that repetitions constitute a form of grammar, such as stress in sentence, for Deaf people (Covington, 1973). A repetition in answer position appears to be part of the answer to the question; thus, ON does not place any emphasis on her answer by repeating it. Second, we observe that ON tends to repeat some expressions in the overall data. It is possible that the repetition is her signing characteristic. We plan to conduct more quantitative analysis comparing other signers in our corpus.

In Analysis 4, we discuss online-communication-specific issues related to the repair sequence in ON's repetition.

## 6.2 Analysis 4: Self-initiated self-repair for a frame-out issue

Figure 8 shows one of the dictionary forms of OKINAWA. In line 02 of Figure 6 and Figure 7, ON's two fingers for answering OKINAWA '(I live in) Okinawa' are frame-out, as shown in Figure 9. Her signing scale is excessively large. In line 04 of Figure 6 and Figure 7, ON reduces her signing scale. This is a successful frame-in, as shown in Figure 10. As the evidence that ON consciously modified her signing scale, after the question-answer adjacency pair, she adjusts the camera position to be captured the upper space of her signing.

This is an example of self-initiated self-repair. In an in-person setting, this type of repair initiation related with frame-out issue does not occur,



右手2指を立て、こめかみからひねるように上へ上げる

Figure 8: An example of dictionary form of OKINAWA (English translation of caption: Hold up the index and middle fingers of the right hand and twist upwards from the temple.) Japanese Federation of the Deaf (2010: 242)



Figure 9: OKINAWA (frame-out, big)



Figure 10: OKINAWA (frame-in, small)

because the signing space is completely opened to between signer and recipients. In online communication, signers monitor how their own signings are viewed by recipients. Occasionally, the signings are frame-out and should be adjusted. This is an online-specific phenomenon.

In the Okinawa view of Figure 7, ON's modification matches as the second pair part (SPP), answering, of the question-answer sequence because SK's SCT in line 03 is delayed. So, line 03 and line 04 are produced almost simultaneously. In the Shikoku view of Figure 6, however, ON's repetition is not placed the second pair part. because SK's SCT in line 03 is not delayed. Because of that, ON's repetition in line 04 floats from the ongoing conversational sequence.

In addition, we notice that some Deaf people tend to increase repetition in online communication more than in-person communication in some small observations of our data-set. At this moment, we plan to compare this type of phenomenon in online and in-person quantitatively for future works.

## 7. Discussion

Levinson (2016) modelled the cognitive mechanisms of turn-taking in everyday human conversation. He estimated intervals of 200 ms to conceptualise one's thoughts, 75 ms to retrieve the lexicon, and 325 ms to encode the form before taking a turn to speak for a total of 600 ms. However, when the timing of turn-taking was measured from actual linguistic data collected worldwide, the start of the response turn was normally distributed with a peak approximately 200 ms after the end of the recipient's turn. He points out that to achieve this, humans plan their own speech production while anticipating their opponent's speech; they also anticipate the end of the turn and follow signals that provide clues to the end of the turn.

Our research question is as follows: What changes would ensue if videoconferencing systems were introduced to the turn-taking process supported by the highly organised human cognitive mechanisms? This is a general question that is common to both spoken dialogues and signed dialogues occurring via videoconferencing systems. Future studies of online communication should consider how recipients accept system-induced latency when basal latency occurs, and how they subsequently interact with each other. Online sign language interaction is an ideal research target to approach this problem because it uses only a video channel without a speech channel.

A limitation of this study is that it is difficult to ascertain whether and how the conversation participants themselves notice and perceive the minute differences in the conversation sequence due to this latency. However, conversation analysis is a research method that analyses how the other party followed the next action in response to a previous action in order to understand the state of awareness of the conversation participants themselves, etc. We will continue to collect data and propose a theory of turn-taking and repair sequences in online communication.

## 8. Conclusion

The technological development of videoconferencing systems, such as Zoom, prioritises the enhancement of usability primarily for hearing people. However, some usability innovations have also been implemented to support the Deaf minority. Although

videoconferencing systems and everyday conversations are not required to be completely equivalent, phenomena including which participant 'greet' the other first or reactions that convey a positive attitude towards the other's utterance, and how the repetitions appear to the remote recipient, as demonstrated in this article, can be significantly inhibited by latency. The sense of accomplishment and satisfaction during a conversation is obtained through a series of interactions with the recipient. We hope that analyses of this nature will be utilised in future efforts to develop video transmission technology.

Thus far, we have merely established the data collection method and data annotation environment. In future studies, we intend to qualitatively and quantitatively analyse the recorded data, then continue the exploration of how Deaf people living in the visual world were forced to confront communicative and cognitive challenges during the COVID-19 pandemic.

## 9. Acknowledgements

We thank our research collaborator; Dr. Ryosaku Makino, who developed the concept of basal latency in online communication with us; Dr. Keiko Sagara, who took part as an interviewer in some sessions; and Prof. Yutaka Osugi, who connected us with the coordinators; the coordinator-in chief, Ms. Megumi Kawakami; area coordinators, Mr. Kazuhiro Naka (Hokkaido), Mr. Ryuji Kondo (Shikoku), and Ms. Eriko Shiroma (Okinawa); and the nine Deaf participants. This study is part of a wider UK–Japanese social science and humanities project seeking to address global challenges presented by the COVID-19 pandemic, with support from the Japan Society for the Promotion of Science (JSPS) International Joint Research Programme JRP-LEAD and UKRI (UK Research and Innovation, UK).

## 10. Bibliographical References

Bono, M., Sakaida, R., Ochiai, K., and Fukushima, S. (2023) Intersubjective Understanding in Finger Braille Interpreter-mediated Interaction: Two Case Studies of Other-initiated Repair. *Lingua*, Elsevier. <https://doi.org/10.1016/j.lingua.2023.103569>

Bono, M., and Adam, R. (2023) Online cross-signing project between the United Kingdom and Japan: First phase of data collection, *Online Proceedings of JSAI-ISA2023*. (published to only conference audience)

Byun K., de Vos, C., Bradford, A., Zeshan, U., and Levinson, S. C. (2018). First encounters: Repair sequences in cross-signing. *Topics in Cognitive Science*, 10(2), 314-334. <https://doi.org/10.1111/tops.12303>

Covington, V. C. (1973). Features of stress in American Sign Language. *Sign Language*

*Studies*, 2, 39–50. <https://doi.org/10.1353/sls.1973.0017>

Dingemans, M. and Enfield, N. (2015). Other-initiated repair across languages: Towards a typology of conversational structures. *Open Linguistics*, 1(1), 96-118. <https://doi.org/10.2478/opli-2014-0007>

Dingemans, M., Kendrick, K. H., Enfield, N. (2016). A coding scheme for other-initiated repair across languages. *Open Linguistics*, 2(1), 35-46. <https://doi.org/10.1515/opli-2016-0002>

Dingemans, M., Torreira, F. and Enfield, N. (2013). Is "Huh?" a Universal Word? Conversational Infrastructure and the Convergent Evolution of Linguistic Items. *PLOS ONE*, 9(4):e94620. <https://doi.org/10.1371/journal.pone.0094620>

ELAN (Version 6.6). (2023). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. <https://archive.mpi.nl/tla/elan>

Floyd, S., Manrique, E., Rossi, G., and Torreira, F. (2016). Timing of visual bodily behavior in repair sequences: Evidence from three languages. *Discourse Processes*, 53(3), 175-204, <https://doi.org/10.1080/0163853X.2014.992680>

Garg, S., Srivastava, A., Glencross, M. and Sharma O. (2022). A study of the effects of network latency on visual task performance in video conferencing. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). Association for Computing Machinery, New York, NY, USA, <https://doi.org/10.1145/3491101.3519678>

Haakana, M., Kurhila, S., Lilja, N., and Savijärvi, M. (2021). Extending sequences of other-initiated repair in Finnish conversation. In Lindström, J., Laury, R., Peräkylä, A., and Sorjonen, M. (Eds.), *Intersubjectivity in Action: Studies in language and social interaction*. John Benjamins, 231-19. ISBN-10: 9027209405, ISBN-13: 978-9027209405

Hayashi, M., Raymond, G., and Sidnell, J. (2013). *Conversational repair and human understanding*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511757464>

Heath, C. and Luff, P.K. (1993). Disembodied conduct: interactional asymmetries in video-mediated communication, G. Button (Ed.), *Technology in Working Order: Studies of Work, Interaction, and Technology*, Rank Xerox Research Centre, London, UK, 35-54.

Hosoma, H. and Muraoka, H. (2022). How can latency in telecommunication affect action sequence analysis? *The Japanese Journal of Language in Society*, 25(1), 230-237. [https://doi.org/10.19024/jails.25.1\\_230](https://doi.org/10.19024/jails.25.1_230)

International Telecommunication Union (ITU), Telecommunication Standardization Sector

- (ITU-T) One-way transmission time. Recommendation G.114, (05/2003).
- Japanese Federation of the Deaf (2010). *Watashi tachi no shuwa: Gakushu jiten (Our sign language: Encyclopedia for learning JSL)*, Japanese Federation of the Deaf Publisher.
- Keating, E., and Mirus, G. (2003). American Sign Language in Virtual Space: Interactions between Deaf Users of Computer-Mediated Video Communication and the Impact of Technology on Language Practices. *Language In Society*, 32(5), 693–714. <http://www.jstor.org/stable/4169299>
- Kendon, A. (2004). *Gesture: Visible Action as Utterance*, Cambridge University Press.
- Kendrick, H., K. (2015). Other-initiated repair in English. *Open Linguistics*, 1, 164-190. <https://doi.org/10.2478/opli-2014-0009>
- Kikuchi, K., and Bono, M. (2013). Sougokoui ni okeru shuwahatuwa wo kijyutu suru tameno anote-shon mojikashuhou no teian (Proposed new annotation and transcription scheme for signed utterances in interaction). *Shuwagaku kenkyuu (Japanese Journal of Sign Language Studies)*, Vol.22, pp.37-63. (written in Japanese) <https://doi.org/10.7877/jasl.22.37>
- Kita, S., van Gijn, I., and van der Hulst, H. (1998). Movement Phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth and M. Fröhlich (Eds.), *Gesture and sign language in human-computer interaction*, International Gesture Workshop Bielefeld, Germany, September 17-19, 1997, *Proceedings. Lecture Notes in Artificial Intelligence* (Vol. 1317, pp. 23-35). Berlin: Springer Verlag.
- Kitzinger, C. (2013). Repair. In J. Sidnell and T. Stivers (Eds.), *The Handbook of Conversation Analysis*, 229–256. NJ: Wiley-Blackwell. <https://doi.org/10.1002/9781118325001>
- Kusters, A. (2015). *Deaf space in Adamorobe: An ethnographic study of a village in Ghana*. Gallaudet University Press.
- Levinson, C. S. (2016). Turn-taking in human communication – Origins and implications for language processing, *Trends in Cognitive Sciences*, 20(1), 6-14. <https://doi.org/10.1016/j.tics.2015.10.010>
- Manrique, E. and Enfield, N. J. (2015). Suspending the next turn as a form of repair initiation: Evidence from Argentine Sign Language. *Frontiers in Psychology*, 15 September 2015 | <https://doi.org/10.3389/fpsyg.2015.01326>
- Manrique, E. (2016). Other-initiated repair in Argentine Sign Language, *Open Linguistics* 2(1), <https://doi.org/10.1515/opli-2016-0001>
- McNeill, D. (1996). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.
- pympi-ling: a Python module for processing ELAN's EAF and Praats TextGrid annotation files. (2013-2021). <https://pypi.python.org/pypi/pympi-ling>
- Schegloff, E. A., Jefferson, G. and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53 (2), 361-382. <https://doi.org/10.2307/413107>
- Schegloff, E. A. (2007). *Sequence Organization in Interaction, A Primer in Conversation Analysis*, 1. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511791208>
- Seuren, L. M., Wherton, J., Greenhalgh, T., and Shaw, S.E. (2021). Whose turn is it anyway? Latency and the organization of turn-taking in video-mediated interaction. *Journal of Pragmatics*, 172, 63-78. <https://doi.org/10.1016/j.pragma.2020.11.005>

## 11. Language Resource References

- Bono, M., and Adam, R. (2023) Online cross-signing project between the United Kingdom and Japan: First phase of data collection, *Online Proceedings of JSAI-isAI2023*. (published to only conference audience)



# Evaluating the Alignment of Utterances in the Swedish Sign Language Corpus

Carl Börstell 

University of Bergen  
Bergen, Norway  
carl.borstell@uib.no

## Abstract

The Swedish Sign Language (STS) Corpus mainly contains segmentations on the lexical level (i.e. signs), which makes it difficult to extract information at clause- or utterance-like levels. In this paper, I evaluate three different methods of segmenting the data into larger units: prosodic, syntactic and translation-based *utterance units*. The results show that none of the utterance units have particularly high accuracy in their alignment with the others, illustrating the challenges facing researchers who are looking to extract meaningful units above the lexical level. In a second step, I extract articulation information from the corpus videos using computer vision methods, but find no clear alignment of articulatory features of the hands and head with the boundaries of the utterance units.

**Keywords:** sign language, corpus, segmentation, clause, utterance, alignment, prosody, computer vision

## 1. Introduction

Today, there is an increasing number of corpora of sign languages in the world (Fenlon and Hochgesang, 2022; Kopf et al., 2022, 2023). Technical approaches can benefit from these resources, as well as facilitate their future expansion (see Morgan et al., 2022). Substantial information can be extracted even from very basic annotations, such as simple lexical level annotations – i.e. segmentations and annotations of each individual sign produced – which tend to be the initial steps of sign language corpora annotation work (Johnston, 2014). While such annotations can provide important insights into, e.g., lexical frequency, collocations and duration (Börstell, 2022b), it is more challenging to use lexical annotations alone to investigate grammatical constructions. This is mainly due to the fact that many sign language corpora lack any form of syntactic segmentation of the signing. One exception is the Auslan Corpus, which features so-called *clause-like units* that internally also have annotations for grammatical functions, enabling more detailed investigations into the syntactic organization of the language (Johnston, 2019). From the perspective of Conversation Analysis, Bono et al. (2020) annotated various layers of linguistic information – e.g., pragmatic, syntactic and phonetic – to segment a corpus of Japanese Sign Language (JSL) dialogues into utterance units based on those combined layers, facilitating research on the interactional aspects of sign language communication.

In this paper, I look at the Swedish Sign Language (STS; *svenskt teckenspråk*) Corpus (Mesch et al., 2012), which does not feature any clause- or utterance-unit segmentations on the whole. However, a small subset of the corpus has previously been annotated for syntactic relations (Östling et al.,

2017), which can be used to infer clause or sentence units for that specific subset. Prosodically motivated segmentation of the corpus has been piloted as well, but was deemed inefficient as a method (Börstell et al., 2014). Without dedicated segmentations above the lexical level, research that required sentence-based segmentations has instead used the translation tier segmentations as an approximation of sentence units (Sjons, 2013; Östling et al., 2015). To date, there has been no evaluation of how past approaches to sentence- or utterance-unit segmentation/approximation align with one another. The goal of this paper is thus to evaluate the equivalence across approximations of utterance units in the STS Corpus, namely those based on available or inferred prosodic, syntactic and translation segmentations.

## 2. Background

The Swedish Sign Language (STS; *svenskt teckenspråk*) Corpus (Mesch et al., 2012) has been available for research since 2011, and has since been published as an online interface (Öqvist et al., 2020). The STS Corpus has mainly been annotated for sign glosses and idiomatic translations into written Swedish (Mesch et al., 2012; Mesch and Wallin, 2015), but has later been enriched with word class annotations (Östling et al., 2015). Smaller subsets have in addition been annotated for other properties such as backchannel responses (Mesch, 2016), mouthings (Mesch et al., 2021) and syntactic segmentations and relations (Börstell et al., 2016). However, there is no comprehensive type of segmentations beyond the original sign and translation tier annotations. In Börstell et al. (2016), we attempted a basic syntactic annotation of the STS Corpus, which involved segmenting clause-



like units on the basis of a combination of syntactic, semantic and prosodic properties of the signing. The definition centered around predicate-type signs as the core, and expressing a single idea within a single prosodic unit, definitions that were further used in later cross-linguistic research (Börstell et al., 2019). In Börstell et al. (2016), the first step was identifying and segmenting a syntactic unit, followed by annotating their internal relations for each sign. This proved to be quite time-consuming, and it involves simultaneous bottom-up and top-down approaches. That is, you need a segmentation to know which signs can relate to each other, but the signs that relate to each other also define the segmentation itself. In several other studies, utterances were inferred on the basis of the translation tier segmentations – i.e. the span of the Swedish translations across signs were used as approximate *utterance units* (defined here as a unit of segmentation corresponding to a level above the sign) – cf. Bono et al. (2020). For example, this was used in approaches to automatically word class tag the STS Corpus (Sjons, 2013; Östling et al., 2015). The translations are, however, not segmented systematically based on the signed articulation, but rather conversational content. In fact, translation annotation was mainly done independently of the sign gloss annotations, based on what could be conveniently expressed in written Swedish. Furthermore, translation segments do not always even correspond to a full sentence in neither Swedish nor STS, as many of them are partial sentences or fragments.

In Börstell et al. (2014), we experimented with ways of segmenting units based on visual prosodic cues, and whether these would correspond to syntactic units. A number of deaf signers were recruited to segment a subset of the STS Corpus based on visual prosodic cues alone, and these were compared to a syntactic segmentation made on the same subset. The results showed a lot of variation in the prosodic segmentations, and whereas some major prosodic breaks aligned across participants, it was deemed less reliable and inefficient as a method for segmenting the corpus data for syntactic purposes. Instead, the work from Börstell et al. (2016) was expanded on later in Östling et al. (2017), when we submitted a subset of the STS Corpus data to the *Universal Dependencies* (De Marneffe et al., 2021) dataset collection, making it the first sign language corpus to be added.<sup>1</sup> There, we instead worked in a bottom-up fashion, annotating grammatical relations between signs individually and later linking them together into a dependency tree automatically, thus skipping

---

<sup>1</sup>STS is the only sign language represented in Universal Dependencies to date, but see Caligiore et al. (2020) for work on Italian Sign Language (LIS).

the explicit segmentation step in the annotation process. The STS dataset in Universal Dependencies is still very small, consisting of 1610 sign glosses across 203 sentences.

Although the Universal Dependencies STS dataset provides syntactic segmentation of clause-like units through its dependency trees, there has not been any evaluation of how well these syntactic units correspond to other units. For instance, to what extent do the syntactic units align with the translation units that have been used as placeholder sentence segmentations in previous work? Would either type of utterance unit, whether syntactic or translation-based, have any meaningful prosodic properties – e.g., notable pauses or other articulatory features around the start-/endpoints. We know from other research that sign language utterances display a multitude of prosodic features that can be used to segment and identify them, such as body, head and eyebrow movements and eyeblinks (Crasborn, 2007; Fenlon et al., 2007; Hansen and Heßmann, 2007; Herrmann, 2010; Sandler et al., 2011; Ormel and Crasborn, 2012; Puupponen et al., 2015; Puupponen, 2019; Kimmelman et al., 2020; Dachkovsky, 2022). Such features have in recent years been used in computer vision-based analyses of sign language data, as part of automatically extracting articulation and potentially segmenting continuous signing (Susman, 2022; Moryossef et al., 2023).

In this paper, I aim to:

1. compare and evaluate the alignment of prosodic, syntactic and translation utterance units in the STS Corpus
2. use computer vision-based tools to investigate articulatory correlates of these units

### 3. Methodology

For this study, I use the six original ELAN (Wittenburg et al., 2006) annotation files (.eaf) used in the annotation of the STS Universal Dependencies dataset (Östling et al., 2017). The six corpus files consist of 12 signers engaged in different types of conversation, between 1.5 and 3 minutes long (14 minutes and 5 seconds in total), comprising 1621 sign tokens: two free conversations (more dialogue) and four stories (more monologue).

The data processing, analysis and visualizations were done in R (R Core Team, 2023) with the packages `ggttext` (Wilke and Wiernik, 2022), `glue` (Hester and Bryan, 2022), `pracma` (Borchers (2022)), `scales` (Wickham and Seidel, 2022), `signglossR` (Börstell, 2022a), `tidyverse` (Wickham et al., 2019) and `udpipe` (Wijffels, 2023). The data and code for this study can be found at: <https://osf.io/fw825/>.

### 3.1. Defining units

The STS data as represented in the Universal Dependencies dataset contains the original sign annotations from the corpus as well as dependency relations between them. These dependency trees form a type of utterance unit segmentation of the STS Corpus data. The utterance units as defined by the Universal Dependencies dependency trees are in the following called *syntactic* utterance units. I compare these *syntactic* utterance units to the so-called *translation* utterance units. The *translation* utterance units are defined as the sign annotations that fall within or overlap with the temporal span of translation tier segmentations. I compare these two unit types also to a third type of utterance unit, labeled *prosodic* utterance units. The *prosodic* utterance units are defined as the sign sequences without any substantial pauses between signs. Here, the pause duration threshold has been set to the median duration of sign pauses between the syntactic units in the Universal Dependencies dataset: 322 milliseconds. Any pause between signs larger than that value forms a segmentation point marking a new prosodic utterance unit. The three types of utterance units – prosodic, syntactic and translation – result in slightly different numbers of utterance units, spanning different numbers of sign annotations (see Table 1).

Unit	# of units	# of signs
Prosodic	264	1621
Syntactic	203	1610
Translation	217	1611

Table 1: The number of utterance units per type and the number of sign annotations covered.

As is visible from Table 1, the largest number of signs is 1621, which is the same as the total number of tokens in the six corpus files of the dataset. This is only found for the prosodic unit segmentation, which is due to the fact that the prosodic segmentation is by definition done on the full dataset of (manual) sign annotations. The translation units have a slightly lower number, because some sign sequences have not been translated (generally short backchannel utterances). The syntactic units have the lowest sign counts because a few sign sequences in the dataset were never annotated for the Universal Dependencies dataset – e.g., due to the annotators being uncertain of the dependency analysis.

While the prosodic and syntactic utterance units always align exactly with the start and end of some sign annotations, since they are defined on the basis of those (sign) annotation segmentations, the translation utterance units do not necessarily align with sign annotation endpoints. Instead, the trans-

lation utterance units are treated as temporal segmentations, which can be aligned to the sign annotations based on overlap: if a sign annotation is completely within the boundaries of a translation unit, it is assigned to it; if a sign annotation overlaps with more than one translation unit, it is assigned to the first overlapping translation unit (see Figure 1).

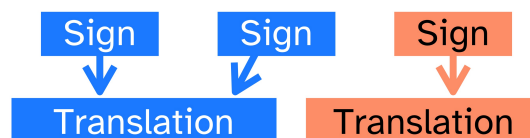


Figure 1: Assignment of signs to translation units.

### 3.2. Measuring Alignment of Units

Alignment across types of utterance units is analyzed in two ways.

First, the content equivalence of segments across tiers is defined as the intersection between unit types with regard to how many identical segments of sign annotations they share. That is, if the sequence of signs ABCDE is segmented as ABC, DE on one tier and A, BC, DE on the other, the two tiers share exactly one segment (i.e. DE).

Second, the temporal alignment and number of segmentations across the types of utterance units are analyzed with the Staccato algorithm (Lücking et al., 2011) as implemented in ELAN (Version 6.2) [Computer software] (2021). The Staccato algorithm is an implementation of the Thomann graph-theoretical method of segment alignment. This method looks at the so-called *degree of organization* of linear segments across tiers, defined as the correspondence of segments into temporally overlapping “shared nuclei” (core overlapping segments).<sup>2</sup> The metric of agreement (*degree of organization*) is based on the amount of overlap as well as the number of identified segments, compared to a chance baseline from iterated Monte Carlo Simulations, thus arriving at a metric between  $-1$  (low) and  $1$  (high), where  $0$  is equal to chance levels in the degree of organization across tiers. Here, the algorithm is run for each pairwise utterance unit tier combination (per file and signer) with 1000 iterations (granularity = 10;  $\alpha = .05$ ). Thus, a value is obtained for every combination of utterance unit segmentation tiers ( $n=30$ ).

### 3.3. Prosody with Computer Vision

Additionally, I extracted articulations through body-pose estimations of the signing in each of the six

<sup>2</sup>See also Rasenberg et al. (2022) for an example of this method used for inter-annotator reliability testing.

corpus files through the computer vision tool *MediaPipe* (Lugaresi et al., 2019). MediaPipe was used to estimate the location of various body landmarks in each of the front-facing videos linked to the ELAN files – thus 12 videos, as there are two signers with one main front-facing video file each for each corpus file. MediaPipe has previously been shown to be successful in analyzing articulatory properties in sign language videos, such as extracting sign articulation onsets and locations (Börstell, 2023) and comparing phonetic features of different text types (Kimmelman and Teresè, 2023).

Here, I focus on the distance moved across frames by 1) the two hands (based on wrist positions in two dimensions) and 2) the head (based on nose position in the vertical dimension), respectively. That is, how far in signing space have the hands and head moved between every sequence of two frames in the video? This is done to identify prosodically prominent points in manual and non-manual articulation – points in time in the files where the hands and/or head move more than usual. The metric used for distance moved is the raw Euclidean distance moved in the MediaPipe coordinate system, but  $z$ -scored within each file and signer for cross-signer and cross-file comparison. The measurements for distance moved by the hands and head were then analyzed for peaks to find sequences of increased activity in relative movement. This was done with the `pracma::findpeaks` function, extracting peaks – defined as frames with a previous increase and following decrease in movement activity ( $\pm 3$  frames) – in the hand and head movement data. With this method, 369 peaks were found in the hand movements across files, and 329 peaks were found in the head movements.

## 4. Results

As seen in Table 1, the syntactic and translation units are more closely overlapping in the total number of units segmented, even though the prosodic unit segmentation was performed on the basis of the median pause duration between syntactic units. When looking at the sign sequences that correspond to each utterance unit (i.e. overlapping sign annotations in the case of translation units), there is a similarity in unit contents that corresponds to the number of units. Table 2 shows the intersection of sign annotation sequence segmentations across utterance unit types, illustrating that the syntactic and translation units have just over 30% overlap in sign sequences resulting from the segmentations, whereas the prosodic utterance units only overlap at around 13–20% with the other utterance unit types. Thus, in terms of content equivalence of sign sequences, it seems the syntactic and trans-

lation segmentations have the highest agreement.

Turning to the general temporal alignment between utterance units, Figure 2 shows all segmentations temporally aligned across the six corpus files. There is, unsurprisingly, agreement on when there is articulation happening in general, but the segmentation endpoints are not always aligned. Although the prosodic utterance units are the most numerous, there are examples where they span much longer stretches of signing than either syntactic or translation units, illustrating sequences with only very short “pauses” between sign annotations. However, we can also see that the translation units are the ones most often entirely mismatched in terms of content, such as including an annotation where the others do not. This happens, for example, by translating non-manual content (e.g., translating visible laughter at the end of file `SSL02_332`) or failing to add a translation annotation in cases of short turns (e.g., several missing annotations in file `SSL01_104` that constitute short response tokens). The missing segments on the syntactic tier are stretches of glosses that are missing from the dependency annotations, thus lacking a corresponding syntactic unit.

As a second type of alignment measure, I used the Staccato algorithm (Lücking et al., 2011) implemented in ELAN to evaluate the agreement between annotation segmentations across utterance unit types. Figure 3 shows the distribution of scores achieved by each comparison, where circles represent each annotation tier comparison and their relative size corresponds to the number of segments per tier (smaller size means fewer segments to match for those tiers). As is visible from Figure 3, the scores obtained in terms of degree of organization are all quite poor, mostly falling at or below chance levels. Opposite to the patterns found for content equivalence in Table 2, the highest scores come from the alignment between prosodic and syntactic units, followed by syntactic and translation units, and lastly prosodic and translation units. Generally, tiers with only a single annotation (usually a single response token or comment by the addressee at the end of a narrative) receive perfect alignment scores, but tiers with many more annotations display much lower agreement.

Turning to the MediaPipe data, Figure 4 shows the movement (distance traveled) of hands and head (solid and dotted lines) within each of the six corpus files. It also shows the major points of segmentation agreement (vertical lines;  $n=89$ ), defined as points in time at which all three utterance unit types have marked the start or end of an annotation segment. The movement data is  $z$ -scored within signers to show relative movement and smoothed with a LOESS function: the solid lines show the articulation of the hands (distance moved by the

		Unit 2 (comparison)		
		Prosodic	Syntactic	Translation
Unit 1	Prosodic	————	41/264; 15.5%	34/264; 12.9%
	Syntactic	41/203; 20.2%	————	65/203; 32.0%
	Translation	34/213; 16.0%	65/213; 30.5%	————

Table 2: The overlap of sign annotation sequences between utterance unit segmentations.

## Alignment of syntactic, translation and prosodic utterance units

Major gaps and discrepancies marked with red circles

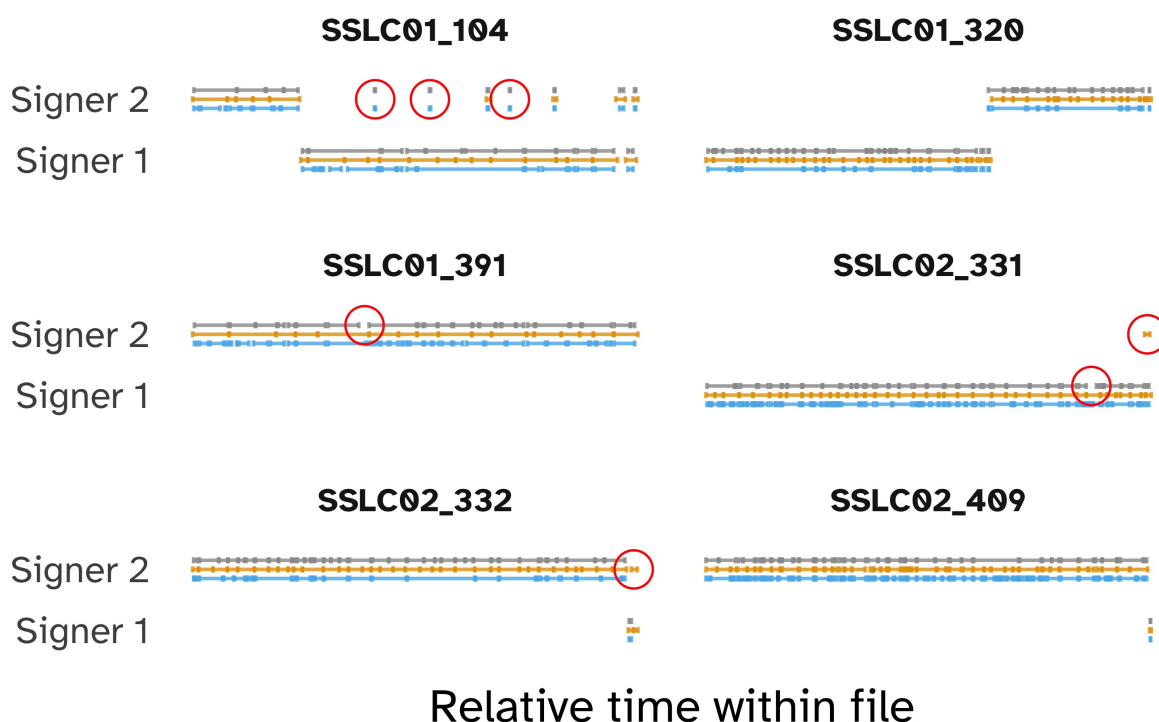


Figure 2: Alignment across utterance units. Red circles mark areas of major discrepancies.

wrist landmarks) and dotted lines show the articulation of the head (vertical distance moved by the nose landmark). The articulation activity can clearly show the main contributor in a text, thus show the major turn-taking events in a conversation (see file SSLC01\_320; NB: Signers with minimal signing in a file have been filtered out here).

Based on Figure 4, there are no obvious visual correlations between the major segmentation points across utterance units and the articulatory activity of the hands and head. Despite some of the segmentation points matching up with either peaks, valleys or changes in overall contour, the picture is too varied to show any obvious patterns of alignment. Out of the identified peaks in the MediaPipe movement data, only 7 (1.9%) of the hand peaks

and 9 (2.7%) of the head peaks occurred within 3 frames of a major segmentation points (i.e. start- or endpoints aligned across all three utterance unit types). Similarly, only 7 (8.1%) and 8 (9.3%) of segmentation points occurred within three frames of a hand or head peak, respectively.

## 5. Discussion and Conclusion

The goal of this study was to evaluate the equivalence and potential usefulness of various types of utterance units in the STS Corpus based on prosodic, syntactic and translation-based segmentations. Seeing as a subset of the STS Corpus is annotated syntactically, these segmentations could



## Degree of organization between utterance units

Distribution of scores for each unit comparison:  
each circle represents one tier comparison

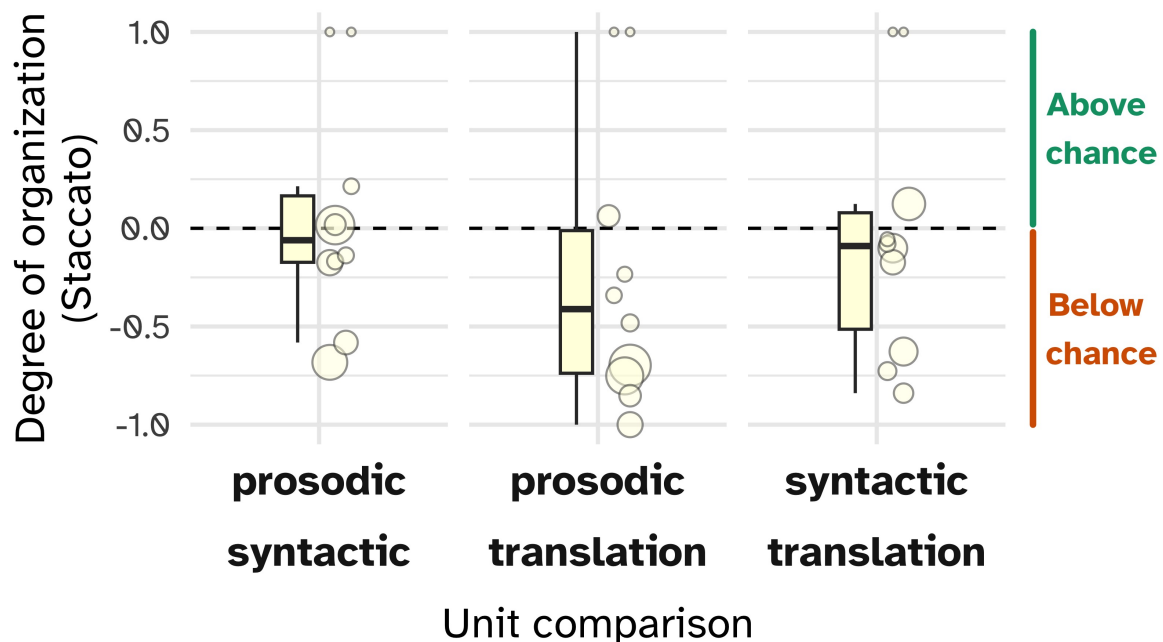


Figure 3: Degree of organization between utterance types using the Staccato algorithm. Circles represent each annotation tier comparison, sizes corresponding to number of annotations per tier. Box plots show the distribution of scores. Dashed lines show chance level.

form a starting point for analyzing the distribution of clause-like units in the corpus, potentially informing automated methods of extracting them. Before such syntactic segmentations were available, the translation tier segmentations had been used as a proxy for a more clause- or sentence-like unit. Segmenting sign annotations into utterance units based on pauses between annotations is another approach, using a type of prosodic (pause duration) information to identify segmentation points.

In this study, it was found that the three methods for identifying utterance units arrive at quite different exact sequences of signs, with at most around 30% overlap in the sequences of signs identified through the different segmentation methods. This shows a low degree of content equivalence between the methods, suggesting that the translation segmentations used in some previous work as a proxy for a sentence-like unit (cf. Sjons, 2013; Östling et al., 2015) do not correspond very closely to the clause-like units identified through manual syntactic annotation (Östling et al., 2017). Nonetheless, the overlap across sign sequence segmentations was higher between syntactic and translation units than any other pairwise comparison. However, the

agreement of segment alignment using the Staccato algorithm (Lücking et al., 2011) pointed to a higher similarity between prosodic and syntactic utterance units than any other pairwise comparison. I suspect this to be the result of the start- and end-points of these units always aligning exactly with sign annotation start- and endpoints, whereas the translation segments are made independently of the sign gloss annotations and rarely align exactly with them at the ends. Additionally, the translation tier segmentations had more instances of complete mismatches compared to the other two tiers, by either adding translations where there were no manual sign annotations or lacking annotations for short manual response tokens (see Figure 2). It is possible that the algorithm is less suitable for this type of data, for which there is often a continuous stream of annotations (i.e. many throughout the file) rather than fewer annotations more sparsely spread out in time. If so, it may not be ideal for evaluating segmentations if the goal of a segmentation is to find the *contents* of what falls within its span, rather than finding its exact endpoints. Another issue is that the number of segments matters for the Staccato algorithm, and the granularity of



## Relative movement of hands and head

Landmark coordinates for wrists (solid) and nose (dotted) of signers **1** & **2**: vertical bars show segmentation points aligned across utterance units

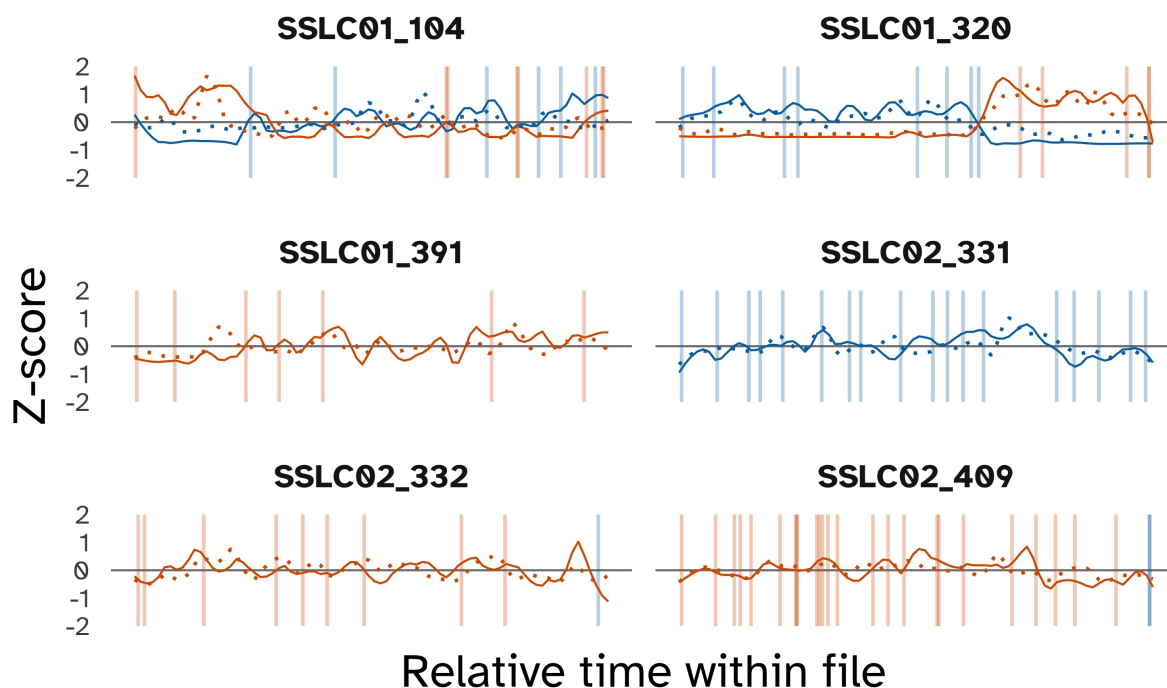


Figure 4: Relative distance moved by hands and head. Solid lines show hand articulation and dotted lines show head articulation (both smoothed with a LOESS function). Vertical lines correspond to utterance unit segmentation points (start or end) matched across all three utterance unit types.

the different methods is quite different as they are based on different motivations: what matters syntactically, what is a convenient content chunk, or what is defined as “pauses”.

The second part of this study looked at prosodic correlates between the identified utterance units and articulatory data extracted from the corpus videos using MediaPipe (Lugaresi et al., 2019). Whereas the extracted data can clearly show patterns such as major turn-taking events between signers in conversation, it was not possible to identify any obvious correlations between shared segmentation points (start or end) across utterance unit types and articulatory patterns in the movement of hands and head. However, seeing as this dataset is only a small subset of the STS Corpus, the lack of found patterns/correlates may simply be due to the lack of sufficient data. A type of hybrid approach was proposed by Chizhikova and Kimmelman (2022), who in their analysis of headshakes and negation used computer vision-based methods together with manual inspection. As the

STS Corpus continues to grow in terms of features annotated for, there will be better opportunities to measure correlations between manually annotated prosodic features and those extracted automatically, as well as using aggregated data from multiple layers of linguistic information – e.g., prosodic, semantic and interactional (cf. Bono et al., 2020) – to arrive at meaningful utterance units.

In summary, this study has shown that the currently available utterance units (whether annotated or inferred) in the STS Corpus do not align to any greater extent. This means that researchers using these units – possibly as a proxy of “sentences” – need to take great care in choosing motivated unit types and be aware of their limitations. The future goal for the STS Corpus should be to segment the sign annotations into some meaningful larger unit, whether conversational turns or utterances or syntactic sentences or clauses. This would increase the potential of the corpus as a language resource substantially, as it would allow for analyses of language structure beyond the individual signs.

## 6. Acknowledgements

I am grateful to Mark Dingemanse for directing me to the Staccato algorithm. I thank three anonymous reviewers for comments and suggestions.

## 7. Data Availability

Data and code are available in an online OSF repository: <https://osf.io/fw825/>

## 8. Bibliographical References

- Mayumi Bono, Rui Sakaida, Tomohiro Okada, and Yusuke Miyao. 2020. [Utterance-Unit Annotation for the JSL Dialogue Corpus: Toward a Multimodal Approach to Corpus Linguistics](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 13–20, Marseille, France. European Language Resources Association (ELRA).
- Hans W. Borchers. 2022. *pracma: Practical Numerical Math Functions*.
- Carl Börstell. 2022a. [Introducing the signglossR Package](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 16–23, Marseille, France. European Language Resources Association (ELRA).
- Carl Börstell. 2022b. [Searching and Utilizing Corpora](#). In Jordan Fenlon and Julie A. Hochgesang, editors, *Signed Language Corpora*, number 25 in Sociolinguistics in deaf communities, pages 90–127. Gallaudet University Press, Washington, DC.
- Carl Börstell. 2023. [Extracting Sign Language Articulation from Videos with MediaPipe](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, NEALT Proceedings Series, No. 52, pages 169–178, Tórshavn, Faroe Islands. University of Tartu Library.
- Carl Börstell, Tommi Jantunen, Vadim Kimmelman, Vanja de Lint, Johanna Mesch, and Marloes Oomen. 2019. [Transitivity prominence within and across modalities](#). *Open Linguistics*, 5:666–689.
- Carl Börstell, Johanna Mesch, and Lars Wallin. 2014. [Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation](#). In *Proceedings of the LREC2014 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, pages 7–10, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Carl Börstell, Mats Wirén, Johanna Mesch, and Moa Gärdenfors. 2016. [Towards an Annotation of Syntactic Structure in the Swedish Sign Language Corpus](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 19–24, Portorož, Slovenia. European Language Resources Association (ELRA).
- Gaia Caligiore, Cristina Bosco, and Alessandro Mazzei. 2020. Building a Treebank in Universal Dependencies for Italian Sign Language. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, pages 1–6, Bologna, Italy. CEUR.
- Anastasia Chizhikova and Vadim Kimmelman. 2022. [Phonetics of Negative Headshake in Russian Sign Language: A Small-Scale Corpus Study](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 29–36, Marseille, France. European Language Resources Association (ELRA).
- Onno Crasborn. 2007. [How to recognise a sentence when you see one](#). *Sign Language & Linguistics*, 10(2):103–111.
- Svetlana Dachkovsky. 2022. [Emergence of a subordinate construction in a sign language: Intonation ploughs the field for morphosyntax](#). *Glossa: a journal of general linguistics*, 7(1).
- Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. [Universal Dependencies](#). *Computational Linguistics*, pages 1–54.
- ELAN (Version 6.2) [Computer software]. 2021. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen. [[link](#)].
- Jordan Fenlon, Tanya Denmark, Ruth Campbell, and Bencie Woll. 2007. [Seeing sentence boundaries](#). *Sign Language & Linguistics*, 10(2):177–200.
- Jordan Fenlon and Julie A. Hochgesang, editors. 2022. *Signed Language Corpora*. Number 25 in Sociolinguistics in deaf communities. Gallaudet University Press, Washington, DC.

- Martje Hansen and Jens Heßmann. 2007. [Matching propositional content and formal markers: Sentence boundaries in a DGS text](#). *Sign Language & Linguistics*, 10(2):145–175.
- Annika Herrmann. 2010. [The interaction of eye blinks and other prosodic cues in German Sign Language](#). *Sign Language & Linguistics*, 13(1):3–39.
- Jim Hester and Jennifer Bryan. 2022. [glue: Interpreted String Literals](#).
- Trevor Johnston. 2014. [The reluctant oracle: Adding value to, and extracting of value from, a signed language corpus through strategic annotations](#). *Corpora*, 9(2):155–189.
- Trevor Johnston. 2019. [Clause constituents, arguments and the question of grammatical relations in Auslan \(Australian Sign Language\): A corpus-based study](#). *Studies in Language*, 43(4):941–996.
- Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. [Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study](#). *PLOS ONE*, 15(6):e0233731.
- Vadim Kimmelman and Anželika Teresė. 2023. [Analyzing literary texts in Lithuanian Sign Language with Computer Vision: a proof of concept](#). In *NAIS 2023: The 2023 symposium of the Norwegian AI Society*, volume 3431 of *CEUR Workshop Proceedings*, page 5, Bergen, Norway. Technical University of Aachen.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2023. [The Sign Language Dataset Compendium](#). Technical report. Version Number: 1.3.
- Andy Lücking, Sebastian Ptock, and Kirsten Bergmann. 2011. [Staccato: Segmentation Agreement Calculator according to Thomann](#). In *Proceedings of the 9th International Gesture Workshop: Gestures in Embodied Communication and Human-Computer Interaction*, pages 50–53, Athens, Greece.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A Framework for Building Perception Pipelines](#). Publisher: arXiv Version Number: 1.
- Johanna Mesch. 2016. [Manual backchannel responses in signers' conversations in Swedish Sign Language](#). *Language & Communication*, 50:22–41.
- Johanna Mesch, Krister Schönström, and Sebastian Embacher. 2021. [Mouthings in Swedish Sign Language: An exploratory study](#). *Grazer Linguistische Studien*, 93:107–135. Publisher: Universität Graz.
- Johanna Mesch and Lars Wallin. 2015. [Gloss annotations in the Swedish Sign Language Corpus](#). *International Journal of Corpus Linguistics*, 20(1):103–121.
- Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. [Sign Language Resources in Sweden: Dictionary and Corpus](#). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, pages 127–130, Paris. European Language Resources Association (ELRA).
- Hope Morgan, Onno Crasborn, Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [Facilitating the Spread of New Sign Language Technologies across Europe](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 144–147, Marseille, France. European Language Resources Association.
- Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. [Linguistically Motivated Sign Language Segmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.
- Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. [STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).
- Ellen Ormel and Onno Crasborn. 2012. [Prosodic Correlates of Sentences in Signed Languages](#):

- A Literature Review and Suggestions for New Types of Studies. *Sign Language Studies*, 12(2):279–315. Publisher: Gallaudet University Press.
- Robert Östling, Carl Börstell, Moa Gårdenfors, and Mats Wirén. 2017. [Universal Dependencies for Swedish Sign Language](#). In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 303–308, Gothenburg, Sweden. Association for Computational Linguistics.
- Robert Östling, Carl Börstell, and Lars Wallin. 2015. [Enriching the Swedish Sign Language Corpus with Part of Speech Tags Using Joint Bayesian Word Alignment and Annotation Transfer](#). In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NoDaLiDa 2015)*, pages 263–268, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.
- Anna Puupponen. 2019. [Towards understanding nonmanually: A semiotic treatment of signers' head movements](#). *Glossa: a journal of general linguistics*, 4(1).
- Anna Puupponen, Tuija Wainio, Birgitta Burger, and Tommi Jantunen. 2015. [Head movements in Finnish Sign Language on the basis of Motion Capture data: A study of the form and function of nods, nodding, head thrusts, and head pulls](#). *Sign Language & Linguistics*, 18(1):41–89.
- R Core Team. 2023. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Marlou Rasenberg, Asli Özyürek, Sara Bögels, and Mark Dingemanse. 2022. [The Primacy of Multimodal Alignment in Converging on Shared Symbols for Novel Referents](#). *Discourse Processes*, 59(3):209–236.
- Wendy Sandler, Irit Meir, Svetlana Dachkovsky, Carol Padden, and Mark Aronoff. 2011. [The emergence of complexity in prosody and syntax](#). *Lingua*, 121(13):2014–2033.
- Johan Sjons. 2013. [Automatic induction of word classes in Swedish Sign Language](#). Master's thesis, Stockholm University.
- Margaux Susman. 2022. [Eye Blinks in French Sign Language Definition of eye blink types and automatic detection of eye blinks using computer vision, rule-based and machine learning-based methods](#). Master's thesis, University of Bergen, Bergen, Norway.
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Pedersen, Evan Miller, Stephan Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. [Welcome to the Tidyverse](#). *Journal of Open Source Software*, 4(43):1686.
- Hadley Wickham and Dana Seidel. 2022. [scales: Scale Functions for Visualization](#).
- Jan Wijffels. 2023. [udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit](#).
- Claus O. Wilke and Brenton M. Wiernik. 2022. [ggtext: Improved Text Rendering Support for 'ggplot2'](#).
- Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. [ELAN: A professional framework for multimodality research](#). In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

## 9. Language Resource References

- Mesch, Johanna and Wallin, Lars and Nilsson, Anna-Lena and Bergman, Brita. 2012. [Dataset. Swedish Sign Language Corpus project 2009–2011 \(version 1\)](#). Sign Language Section, Department of Linguistics, Stockholm University. PID <https://hdl.handle.net/1839/b9b9c88a-f8df-4fa5-8eb0-53622108764d>.



# How to Approach Lexical Variation in Sign Language Corpora

Carl Börstell 

University of Bergen  
Bergen, Norway  
carl.borstell@uib.no

## Abstract

Looking at lexical frequency and, by extension, lexical variation is often among the first objectives after compiling a sign language corpus, since the only prerequisite is existing sign gloss annotations. However, measuring lexical frequency in a theoretically and statistically meaningful way can be a challenge. In this paper, I provide an overview of how to approach lexical variation in sign language corpora. The aim is to show ways of tackle lexical variation from different angles, from data collection to statistics and visualization, and how to motivate choices based on the data available and the research goals, thus serving as a practical guide for sign language corpus research. Drawing from previous work by different sign language corpus project teams, various approaches to measuring lexical variation are illustrated with data from the Swedish Sign Language (STS) Corpus, with examples that can easily be adapted to any sign language corpus.

**Keywords:** sign language, corpus, lexical frequency, variation, sociolinguistics

## 1. Introduction

The number of available sign language corpora in the world is constantly increasing, and many corpora of individual sign languages are also growing in size (see, e.g., Kopf et al., 2021, 2022, 2023; Fenlon and Hochgesang, 2022). The first step of annotating a sign language corpora is often to segment and annotate individual *signs* in the data (Johnston, 2010). With annotation of individual lexical items (i.e. *signs*), an easy first exploration of the corpus data is to look at lexical frequencies – which signs are used the most, by whom and in what context? Lexical frequency has been studied for a number of sign languages already, with datasets of varying size (e.g., Morford and MacFarlane, 2003; McKee and Kennedy, 2006; Johnston, 2012; Fenlon et al., 2014; Börstell et al., 2016).

It is well known that the distribution of words in language(s) is extremely skewed, with a small number of words occurring frequently but most words occurring fairly rarely (Zipf, 1935). This skew in token frequencies needs to be taken into account when looking at lexical frequency, and makes it more challenging to look at lexical variation, especially in smaller corpora – and most sign language corpora are still relatively small. Thus, there are several aspects to consider when investigating lexical variation within individual sign languages, and I will in the following provide concrete examples of approaches taken in previous work, and opportunities and issues that come with them. While mostly illustrated with examples from the Swedish Sign Language (STS; *svenskt teckenspråk*) Corpus (Öqvist et al., 2020), the methods could be applied to any sign language corpus. Finally, the paper concludes with a summarized list of benefits and downsides to different approaches and metrics.

## 2. Data and Methods

For the examples in this paper, I use data from the STS Corpus (Öqvist et al., 2020) presented in different ways depending on the approach to investigating lexical variation.

The STS Corpus data (Mesch et al., 2012) was retrieved from *The Language Archive* (<https://archive.mpi.nl/tla/>) in July 2023 and consists of 189,679 sign tokens across 298 annotation files and 42 signers.

The data was retrieved, processed and visualized using R v4.3.2 (R Core Team, 2023) and the packages `patchwork` v1.2.2 (Pedersen, 2022), `scales` v1.2.1 (Wickham and Seidel, 2022), `sign-glossR` v2.2.4 (Börstell, 2022), `tidylo` v0.2.0 (Schnoebelen et al., 2022) and `tidyverse` v2.0.0 (Wickham et al., 2019).

Simulated example data and code for calculating and plotting frequencies and variation can be found at: [https://github.com/borstell/r\\_functions/blob/main/plotting\\_corpus\\_variation.R](https://github.com/borstell/r_functions/blob/main/plotting_corpus_variation.R)

## 3. Approaches to Lexical Variation

In order to look at lexical variation in any language, one needs to have enough data, such that it covers the relevant variables involved in variation – whether, e.g., age, gender or geographic belonging (Bayley et al., 2015). While variation can be studied separately from a corpus, through interviews and elicitation with the signing community directly (Lucas et al., 2009; Fisher et al., 2016; Saifar, 2021) or indirectly through distributed surveys online (Kimmelman et al., 2022), the focus in this paper is data collected within a sign language corpus project. However, even within corpus projects,



similar alternative data collection approaches have been used. For example, several projects have included a targeted lexical elicitation task as part of the corpus data collection – i.e. tasks alongside the collection of naturalistic conversational data. The targeted interview/elicitation approach facilitates comparisons of signs in domains known for variation, such as color terms in British Sign Language (BSL) (Stamp et al., 2014) and German Sign Language (DGS) (Langer, 2012), as it results in a larger target sample. Some corpus projects have also adopted a method of crowdsourcing signs and lexical variation as well as perceptions about variation and usage of already documented variants through direct or online community involvement (Kankkonen et al., 2018; Wähl et al., 2018; Hanke et al., 2020). Targeted elicitation tasks are suitable for comparing variation between different groups with regard to specific items/domains since it results in a higher number of data points per item and a better coverage with many signers being represented (cf. Section 3.4). However, elicited data will not be directly comparable to other items/domains found only in the conversational portion of the corpus data, as the distribution of occurrences will look very different.

In the following sections, I will mainly focus on how to approach and measure lexical variation in naturalistic, conversational corpus data.

### 3.1. Counts: “How Many Have You Got?”

As was mentioned in the introduction, the Zipfian distribution of lexical items in a corpus means that token frequencies will be extremely skewed: some items are very frequent whereas most items are very infrequent. Thus, raw counts of frequencies are often quite uninformative as they are only meaningful for a particular corpus (or, corpus size) and will have a huge range between items in the upper vs. lower end of the frequency span. For example, saying that there are 10,846 occurrences of PRO1 (first-person pronoun), 414 occurrences of TYP@b (‘kinda’; fingerspelled) and 7 occurrences of ÄLG(Jbt) (‘moose’) in the STS Corpus is quite meaningless unless they are compared to the total number of tokens in the corpus (n=189,679) or possibly to each other. Nonetheless, in the online STS Dictionary (teckenspråkslexikon, 2023), the only currently available information about corpus frequencies of dictionary entries is raw corpus frequencies, available for those entries that have been linked to the corpus (cf. Mesch et al., 2012). This was why we in Börstell and Östling (2016) developed a search tool for exploring meaningful lexical frequencies and variation in the STS Corpus by rather focusing on *relative* frequencies within and across groups of signers or text types, which is discussed further in Section 3.2.

### 3.2. Proportions: “It’s All Relative!”

One way of approaching *relative frequencies* in a corpus is to simply say how many times an item occurs relative to the total, usually rescaled to arrive at a more interpretable number, e.g., occurrences per 100,000 tokens. This means that we could reformulate the frequencies in Section 3.1 and say that PRO1 occurs 5,718 times per 100,000 tokens, TYP@b 218 times per 100,000 tokens and ÄLG(Jbt) about 4 times per 100,000 tokens. This metric is more intuitive and more useful as it is comparable across corpora or subcorpora of different sizes. However, it does not address the issue of variation, as it does not differentiate where the tokens come from within the corpus.

In Börstell and Östling (2016), we identified the need to obtain relative frequencies of signs in the STS Corpus with attention to sociolinguistic variation. Thus, we developed an online search tool<sup>1</sup>, parallel to the STS Corpus, that would display relative frequencies within different grouping variables that were likely to exhibit variation in lexical frequency distribution: age, gender, region and text type. Thus, frequencies were relative to the total number of tokens by subgroup. This allowed for comparisons across groups of different sociolinguistic variables very easily. For example, there was anecdotal evidence of the sign TYP@b (‘kinda’; fingerspelled) being more frequent among younger signers, and this was corroborated with our search tool illustrating relative frequencies, showing that the sign is much more frequent among younger age groups. Figure 1 shows the same pattern in the current version of the STS Corpus, with over twice the number of tokens annotated compared to what was reported in Börstell and Östling (2016).

One potential feature that was not available in the search tool by Börstell and Östling (2016) was directly comparing relative proportions between multiple forms for the same meaning. Many sign languages exhibit variation in specific domains (e.g., numerals and color terms), such that the same meaning may be expressed by multiple forms. Such variation may consist of either completely different lexical items or phonological variants of a similar base (or iconic mapping), sometimes with sociolectal differences in their distribution (see, e.g., McKee et al., 2011; Langer, 2012; Stamp et al., 2014; Wähl et al., 2018; Safar, 2021; Lutzenberger et al., 2021, 2023). A rather straightforward way of comparing differences in the distribution of sign variants for the same meaning is to compare the

---

<sup>1</sup>The tool, *SSL-lects*, has been offline for a few years due to server replacements and anonymization concerns with the raw STS Corpus data, but there have been plans to integrate a similar tool directly in the online corpus and dictionary resources.

## Relative token frequencies for TYP@b

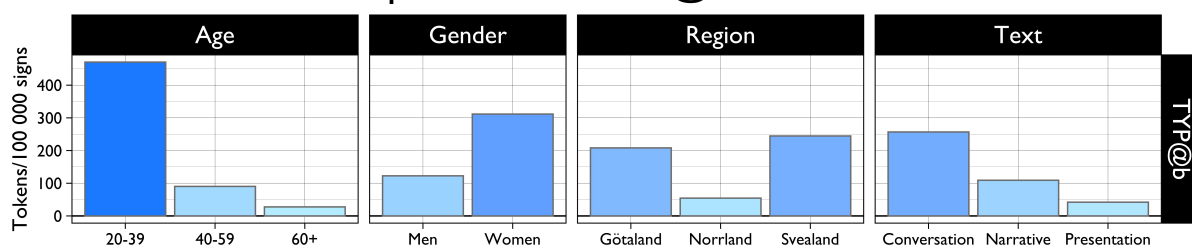


Figure 1: Relative frequencies of the sign TYP@b ('kinda'; fingerspelled) across sociolinguistic groupings in the STS Corpus.

proportion of tokens they each have relative to their combined total, distributed across the sociolinguistic groupings of interest. For example, Figure 2 shows the relative proportions between a one- and two-handed (phonological) variant of the sign for '(an)other' in STS. Based on the relative proportions alone, it is quite clear that the one-handed variant is more common overall but that the oldest signers have a slight preference for the two-handed variant.

Searching for lexical variants or any signs with related meanings is, however, not necessarily straightforward. Glosses are often selected on the basis of a written word with similar meaning, but semantic extension and polysemy may mean that signs are related without sharing a similar gloss (cf. Johnston, 2010; Ormel et al., 2010). Because of this, searching for variants or related signs may already require some knowledge about the language as well as the annotation conventions of the corpus (e.g., how glosses are used).<sup>2</sup>

With these approaches, one issue is that they mainly target specific signs (individually or paired) that we already suspect may display some type of sociolectal variation in their distribution. In Section 3.3, we will see how other metrics can be used to identify interesting distributional variation directly from the data.

### 3.3. Ratio: “What Are the Odds?”

Looking at frequencies relative to sociolinguistic groupings made it possible to visualize variation differences for items suspected to exhibit variation. However, in Börstell and Östling (2016), we also wanted to find ways of *identifying* potential variation-exhibiting items without necessarily knowing about them through previous – often anecdotal – evidence. Thus, we applied a Bayes factor approach, calculating distributions relative to token counts among the same sociolinguistic groupings and could identify certain signs that were overrepresented in some subgroup. While this metric was not

<sup>2</sup>I thank a reviewer for raising this point.

available in the search and visualization tool itself, it could be an interesting addition since it is possible to see both positive and negative values, and as such the directionality of frequency: higher or lower than expected. In Figure 3, a similar implementation is used in a visualization, but with weighted log odds using a Bayesian prior estimated from the data itself, which accounts for differences in sampling variability (see Monroe et al., 2008; Schnoebelen et al., 2022). With this approach, we can confirm that age is a major factor in the distribution of tokens, with TYP@b being skewed towards younger age groups. The gender distribution here is less informative, seeing as the STS Corpus has more women in the younger age groups and more men in the older age groups. Somewhat surprisingly, the text type distribution in Figure 3 is switched compared to Figure 1, which is a consequence of the informative prior taking the sampling variability into account – using an uninformative prior will instead correspond more closely to the relative frequencies in Figure 1, albeit on a different scale.

A log odds approach was also taken by Stamp et al. (2014), who looked at larger groups of signs in specific domains (e.g., numerals and color terms) to see differences in the use of traditional (often regional) signs for concepts in these domains, finding that age was an important factor, with older signers being more likely to use the traditional signs with regional variation, while younger signers exhibit less variation, pointing to dialectal leveling.

### 3.4. Spread & Coverage: “The One with All the Tokens”

As has been mentioned earlier, lexical variation in corpus data can be a challenge due to the low token frequency of most lexical items even in large corpora, which means it is difficult to find items that occur across, e.g., sociolinguistic groupings in spontaneous, conversational data. This is why several corpus projects have opted to include an explicit lexical elicitation task as part of the data collection – this is, however, not the case for the STS

## Relative token frequencies for ANNAN(ea) and ANNAN(ml)

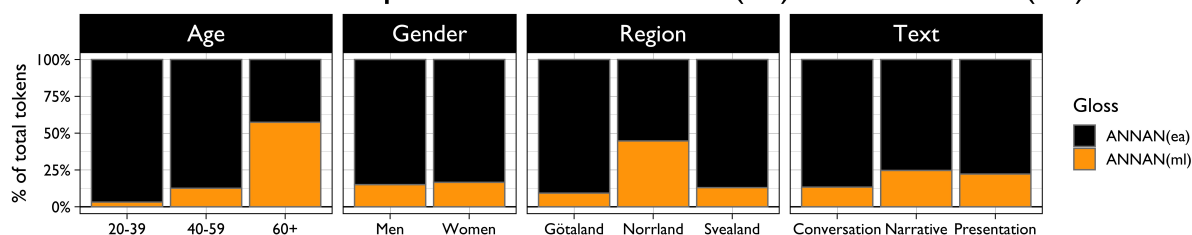


Figure 2: Relative proportions of the signs ANNAN(ea) ('(an)other'; one-handed) and ANNAN(ml) ('(an)other'; two-handed) across sociolinguistic groupings in the STS Corpus.

## Weighted log odds for TYP@b

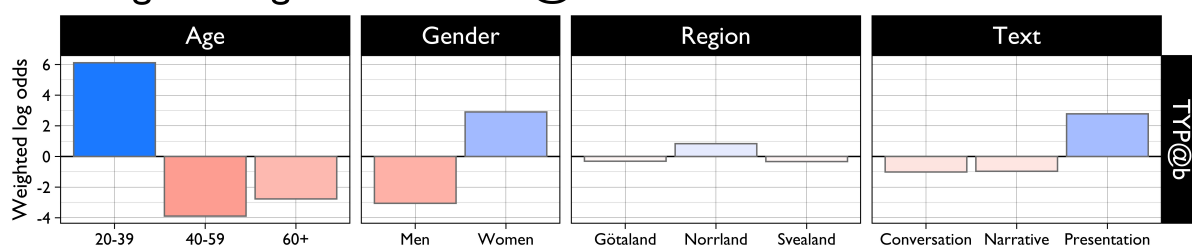


Figure 3: Weighted log odds of the sign TYP@b ('kinda'; fingerspelled) across sociolinguistic groupings in the STS Corpus.

Corpus. It also means that any grouped metric, such as relative frequencies per age group, should also include a measure of spread across signers, at least for low-frequency items – that is, how many signers in the data use the sign at least once (i.e. signer coverage). As an example, in [Börstell and Östling \(2016\)](#) we discussed the known regional variation between two signs for 'moose' in STS: one that depicts the horns (considered the more general and widespread sign) and one that depicts the snout/muzzle (considered a northern variant). In our paper, we noticed that only the “northern” variant was present in the data, found in the northern (Norrland) region as expected. However, not only is it impossible to establish the source of variation, due to the lack of tokens for the other variant, the signer coverage was very poor, with all occurrences being produced by a single signer. In the current, larger STS Corpus dataset, the pattern is unfortunately still the same, with only one of the two variants being produced with 7 occurrences in the whole corpus, all produced by the same signer: an older man from Norrland. Since it is clearly impossible to generalize from a single signer, it can be wise to include signer coverage in a visualization or simply checking the distribution across signers when looking at any token frequencies, but particularly lower ones. Figure 4 shows an example of the signer coverage for three signs, PRO1, TYP@b and ÄLG(Jbt), with dots representing each of the 42

signers in the STS Corpus, where the blue ones represent signers with attested tokens (darker means a higher proportion of total tokens) and grey ones represent signers without attested tokens. As this figure shows, highly frequent signs such as PRO1 will have a large and fairly even spread across signers, whereas signs such as ÄLG(Jbt) cannot be generalized in their usage despite having more occurrences ( $n=7$ ) than the global median number of tokens ( $n=1$ ) in the whole corpus.

### 3.5. Topics & Representativeness: “What Are We Talking About?”

Small(er) corpora, such as most sign language corpora, are quite susceptible to idiosyncrasies skewing the data. For example, multiple sign language corpora have included the same elicitation tasks to elicit narrative texts. Because of this, it comes as no surprise that signs for concepts such as ‘snowman’ and ‘frog’ may be much more frequent than expected from any regular conversation within the deaf community, simply due to the influence of the contents in the elicitation stimuli. Specific topics, and consequently associated words/signs, will always be subject to sampling procedures in the data collection, regardless of the type of corpus. Since sign language corpora involve members of the deaf or signing community, it is expected that concepts such as ‘deaf’ and ‘hard-of-hearing’ may be orders

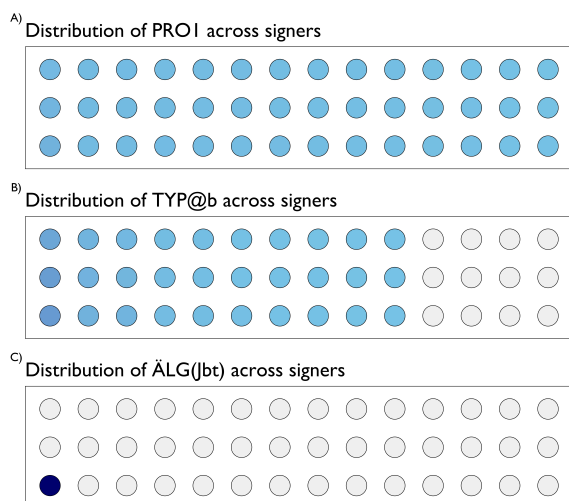


Figure 4: Distribution of tokens across signers for three signs : A) PRO1; B) TYP@b; C) ÄLG(Jbt). Each dot represents a signer; blue-filled dots show signers with attested tokens, with the darkness of the fill color representing proportion of total tokens.

of magnitude more frequent in a sign language corpus than any spoken language corpora. This is not a problem as it directly reflects themes and topics that are relevant in the community, but other topics that are introduced due to targeted tasks in the data collection procedure will often result in some lexical items being overrepresented in a way that is not representative of issues of particular significance to the community at large.

While the use of similar topics/content across sign language corpora is a great resource for cross-linguistic work on, e.g., grammatical and discourse structure (cf. Ferrara et al., 2022), it inadvertently leads to a skew in particular lexical items, which should be taken into account when looking at lexical frequency and variation.

### 3.6. Conventions & Conventionalization: “That’s Not Even a Word!”

As discussed in more detail by Langer et al. (2016), not all tokens are necessarily representative of the regular usage of the individual signer who produced them. For example, some signs are used metalinguistically, in the sense that sign variants are produced i) to illustrate how *others* sign something, ii) as a direct copy of the interlocutor’s sign choice, or iii) to emphasize how the signer themselves does *not* sign (Langer et al., 2016, 140). Similarly, signs may also be produced in a manner different from established lexical items in the language, such as being produced in a context showing, e.g., how

non-signers or learners are attempting to sign or gesture (Langer et al., 2016, 141).

Furthermore, Langer et al. (2016, 141) also mention slips of the hand (i.e. errors in producing the target sign form). This is a question that very much concerns the annotation process in building a corpus, whether to mark accidental deviations/errors explicitly or to simply annotate target forms (if identifiable). In the Auslan Corpus, the procedure for fingerspelling has been to annotate both target form and actual realization in the same sign gloss (Johnston, 2019, 45). This way, the researcher could choose whether to focus on target forms or actual realization, which in itself would be relevant for lexical variation. In the STS Corpus, uncertain or interrupted glosses have been marked with special tags (“@z” and “@&”, respectively), but there is also a dedicated tag for so-called *home-made signs* (“@hg”), which are not considered established signs of the community as a whole (Mesch and Wallin, 2021, 25–26). While such signs make great candidates for a detailed analysis of lexical variation, they will not be generalizable to the larger community. Thus, a researcher interested in investigating lexical variation would need to know the annotation conventions of the specific corpus to be able to accurately match sign glosses to actual forms, and to motivate their reasons for including or excluding specific items.

## 4. Discussion & Conclusions

In this paper, I have given a brief introduction to the question of how to approach lexical variation in sign language corpora. The goal has been to provide anyone interested in doing research on a sign language corpus with concrete examples of issues to consider both theoretically and practically. How the data is annotated will directly influence what can be researched, and which analysis method is applied will affect the usefulness and interpretation of the results. For example, can related signs (e.g., lexical variants) be matched and compared based on glosses alone? Can glosses and search patterns easily distinguish phonological from lexical variants of the same meaning? Are we able to search lemma forms but still account for the frequency of different morphological forms (e.g., inflections) of that lemma? Can we easily attribute tokens to individual signers, and group signers and files by metadata features? These issues are concerns of the researcher using and searching the corpus as much as of the developer of the corpus resource itself, and require users to be familiar with both the language and the corpus conventions.

Unfortunately, few sign language corpora have integrated tools for directly querying a database and receiving a table or visualization of the search re-



sults in a meaningful way, such as regional variation visualized on a map (however, see Hanke, 2016; Hanke et al., 2023). Since lexical variation is an important part in applied areas such as language teaching and interpreting, it would be useful to incorporate simple search tools into the sign language corpus resources – see Isard and Konrad (2022) and Isard and Konrad (2023). Such tools could display not only raw search hits of sign glosses, but also relevant summaries of results presented as tables, graphs or maps, based on variables and metrics selected by the user. In the case of the STS Corpus (Öqvist et al., 2020), the current on-line interface with streamed videos and glosses is a great resource for teachers and students, but it unfortunately does not allow the user to query the database about relative frequencies or proportions between variants, nor export raw search results to be investigated externally, which renders it less accessible to the corpus linguist.

For the researcher who wants to approach questions of lexical frequency and variation in a sign language corpus, here are some points to consider when retrieving, interpreting and reporting the results:

- **Raw frequency:** Numbers will naturally be very skewed due to the Zipfian distribution of lexical items in any corpus and language. Logarithmic scaling can help for visualization purposes.
- **Relative frequency:** Metrics such as *occurrences per 100,000 tokens* will be more useful for comparisons across corpora/languages than raw frequencies, but will nonetheless be skewed across lexical items (i.e. *signs*).
- **Relative proportion:** A useful metric when comparing lexical or phonological variants for the same meaning, but will often suffer from a lack of data unless targeted lexical elicitation was part of the data collection.
- **Log odds:** Log odds are useful to show differences in frequency distributions based on some grouping variable (e.g., gender, region, text type) by accounting for imbalances in raw frequencies for different items, but will not distinguish form variation from differences in conversational content (i.e. *topics*). Note that the weighting and priors used will impact the results, so choose a method that suits your purposes.
- **Signer coverage:** Group-based variation (e.g., gender or region) in corpus data should preferably also account for signer coverage to ensure that the usage reflects the group as a whole rather than a single individual (signer) within it.
- **Type of usage:** Some items may be used incorrectly (e.g., slip of the hand) or metalinguistically (e.g., commenting on how *others* sign (see Langer et al., 2016), and it is thus important to investigate how and why individual items occur in a specific context – especially for low-frequency items.
- **Annotation conventions:** Know the annotation conventions of the corpus you are using, as this directly impacts both what questions you can ask with the data and how to interpret the results.

## 5. Bibliographical References

- Robert Bayley, Adam C. Schembri, and Ceil Lucas. 2015. *Variation and change in sign languages*. In Adam C. Schembri and Ceil Lucas, editors, *Sociolinguistics and Deaf Communities*, 1 edition, pages 61–94. Cambridge University Press, Cambridge.
- Carl Börstell. 2022. *Introducing the signglossR Package*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 16–23, Marseille, France. European Language Resources Association (ELRA).
- Carl Börstell, Thomas Hörberg, and Robert Östling. 2016. *Distribution and duration of signs and parts of speech in Swedish Sign Language*. *Sign Language & Linguistics*, 19(2):143–196.
- Carl Börstell and Robert Östling. 2016. *Visualizing Lects in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language*. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 13–18, Portorož, Slovenia. European Language Resources Association (ELRA).
- Jordan Fenlon and Julie A. Hochgesang, editors. 2022. *Signed Language Corpora*. Number 25 in *Sociolinguistics in deaf communities*. Gallaudet University Press, Washington, DC.
- Jordan Fenlon, Adam Schembri, Ramas Rentelis, David Vinson, and Kearsy Cormier. 2014. *Using conversational data to determine lexical frequency in British Sign Language: The influence of text type*. *Lingua*, 143:187–202.
- Lindsay Ferrara, Benjamin Anible, Gabrielle Hodge, Tommi Jantunen, Lorraine Leeson, Johanna



- Mesch, and Anna-Lena Nilsson. 2022. [A cross-linguistic comparison of reference across five signed languages](#). *Linguistic Typology*, 0(0).
- Jami N. Fisher, Julie A. Hochgesang, and Meredith Tamminga. 2016. [Examining Variation in the Absence of a 'Main' ASL Corpus: The Case of the Philadelphia Signs Project](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 75–80, Portorož, Slovenia. European Language Resources Association (ELRA).
- Thomas Hanke. 2016. [Towards a Visual Sign Language Corpus Linguistics](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 89–92, Portorož, Slovenia. European Language Resources Association (ELRA).
- Thomas Hanke, Elena Jahn, Sabrina Wähl, Oliver Böse, and Lutz König. 2020. [SignHunter – A Sign Elicitation Tool Suitable for Deaf Events](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 83–88, Marseille, France. European Language Resources Association (ELRA).
- Thomas Hanke, Reiner Konrad, and Gabriele Langer. 2023. [Exploring regional variation in the DGS Corpus](#). In Ella Wehrmeyer, editor, *Studies in Corpus Linguistics*, volume 108, pages 192–218. John Benjamins Publishing Company, Amsterdam.
- Amy Isard and Reiner Konrad. 2022. [MY DGS – ANNIS: ANNIS and the Public DGS Corpus](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association.
- Trevor Johnston. 2010. [From archive to corpus: Transcription and annotation in the creation of signed language corpora](#). *International Journal of Corpus Linguistics*, 15(1):106–131.
- Trevor Johnston. 2012. [Lexical frequency in sign languages](#). *Journal of Deaf Studies and Deaf Education*, 17(2):163–193.
- Trevor Johnston. 2019. [Auslan Corpus Annotation Guidelines](#).
- Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch, and Carl Börstell. 2018. [Crowdsourcing for the Swedish Sign Language Dictionary](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 171–176, Miyazaki, Japan. European Language Resources Association (ELRA).
- Vadim Kimmelman, Anna Komarova, Lyudmila Luchkova, Valeria Vinogradova, and Oksana Alekseeva. 2022. [Exploring Networks of Lexical Variation in Russian Sign Language](#). *Frontiers in Psychology*, 12:740734.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. [Overview of Datasets for the Sign Languages of Europe](#). Publisher: Universität Hamburg Version Number: 1.0.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2023. [The Sign Language Dataset Compendium](#). Technical report. Version Number: 1.3.
- Gabriele Langer. 2012. [A colorful first glance at data on regional variation extracted from the DGS-Corpus: With a focus on procedures](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 101–108, Istanbul, Turkey. European Language Resources Association (ELRA).
- Gabriele Langer, Thomas Hanke, Reiner Konrad, and Susanne König. 2016. [“Non-tokens”: When Tokens Should not Count as Evidence of Sign Use](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 137–142, Portorož, Slovenia. European Language Resources Association (ELRA).
- Ceil Lucas, Robert Bayley, and Clayton Valli. 2009. [Sociolinguistic Variation in American Sign Language](#). Gallaudet University Press.
- Hannah Lutzenberger, Connie de Vos, Onno Crasborn, and Paula Fikkert. 2021. [Formal variation in the Kata Kolok lexicon](#). *Glossa: a journal of general linguistics*, 6(1).
- Hannah Lutzenberger, Katie Mudd, Rose Stamp, and Adam Charles Schembri. 2023. [The social structure of signing communities and lexical variation: A cross-linguistic comparison of three unrelated sign languages](#). *Glossa: a journal of general linguistics*, 8(1).

- David McKee and Graeme Kennedy. 2006. [The distribution of signs in New Zealand Sign Language](#). *Sign Language Studies*, 6(4):372–391.
- David McKee, Rachel McKee, and George Major. 2011. [Numeral Variation in New Zealand Sign Language](#). *Sign Language Studies*, 12(1):72–97. Publisher: Gallaudet University Press.
- Johanna Mesch and Lars Wallin. 2021. [Annoteringskonventioner för teckenspråkstexter. Version 8. \[Annotation guidelines for sign language texts\]](#).
- Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. [Sign Language Resources in Sweden: Dictionary and Corpus](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 127–130, Istanbul, Turkey. European Language Resources Association (ELRA).
- Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. [Fightin’ Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict](#). *Political Analysis*, 16(4):372–403.
- Jill P. Morford and James MacFarlane. 2003. [Frequency characteristics of American Sign Language](#). *Sign Language Studies*, 3(2):213–226.
- Ellen Ormel, Onno Crasborn, Els van der Kooij, Lianne van Dijken, Ellen Yassine Nauta, Jens Forster, and Daniel Stein. 2010. [Glossing a multi-purpose sign language corpus](#). In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 186–191, Valletta, Malta. European Language Resources Association (ELRA).
- Thomas Lin Pedersen. 2022. [patchwork: The Composer of Plots](#).
- R Core Team. 2023. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Josefina Safar. 2021. [What’s your sign for TOR-TILLA? Documenting lexical variation in Yucatec Maya Sign Languages](#). *Language Documentation & Conservation*, 15:30–74.
- Tyler Schnoebelen, Julia Silge, and Alex Hayes. 2022. [tidylo: Weighted Tidy Log Odds Ratio](#).
- Rose Stamp, Adam Schembri, Jordan Fenlon, Ramas Rentelis, Bencie Woll, and Kearsy Cormier. 2014. [Lexical variation and change in British Sign Language](#). *PLoS ONE*, 9(4).
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. [Welcome to the tidyverse](#). *Journal of Open Source Software*, 4(43):1686.
- Hadley Wickham and Dana Seidel. 2022. [scales: Scale Functions for Visualization](#).
- Sabrina Wähl, Gabriele Langer, and Anke Müller. 2018. [Hand in Hand - Using Data from an Online Survey System to Support Lexicographic Work](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 199–206, Miyazaki, Japan. European Language Resources Association (ELRA).
- George K. Zipf. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin, New York, NY.
- Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. [STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).

## 6. Language Resource References

- Isard, Amy and Konrad, Reiner. 2023. [MY DGS – ANNIS](#). Hamburg University.
- Mesch, Johanna and Wallin, Lars and Nilsson, Anna-Lena and Bergman, Brita. 2012. [Dataset. Swedish Sign Language Corpus project 2009–2011 \(version 1\)](#). Sign Language Section, Department of Linguistics, Stockholm University. PID <https://hdl.handle.net/1839/b9b9c88a-f8df-4fa5-8eb0-53622108764d>.
- Svenskt teckenspråkslexikon. 2023. [Swedish Sign Language Dictionary online](#). Dept. of Linguistics, Stockholm University.

# Systemic Biases in Sign Language AI Research: A Deaf-Led Call to Reevaluate Research Agendas

Aashaka Desai , Maartje De Meulder , Julie A. Hochgesang ,  
Annemarie Kocab , and Alex X. Lu 

University of Washington, University of Applied Sciences Utrecht/Heriot-Watt University,  
Gallaudet University, Johns Hopkins University, Microsoft Research  
aashakad@cs.washington.edu, maartje.demeulder@hu.nl, julie.hochgesang@gallaudet.edu,  
kocab@jhu.edu, lualex@microsoft.com

## Abstract

Growing research in sign language recognition, generation, and translation AI has been accompanied by calls for ethical development of such technologies. While these works are crucial to helping individual researchers do better, there is a notable lack of discussion of systemic biases or analysis of rhetoric that shape the research questions and methods in the field, especially as it remains dominated by hearing non-signing researchers. Therefore, we conduct a systematic review of 101 recent papers in sign language AI. Our analysis identifies significant biases in the current state of sign language AI research, including an overfocus on addressing perceived communication barriers, a lack of use of representative datasets, use of annotations lacking linguistic foundations, and development of methods that build on flawed models. We take the position that the field lacks meaningful input from Deaf stakeholders, and is instead driven by what decisions are the most convenient or perceived as important to hearing researchers. We end with a call to action: the field must make space for Deaf researchers to *lead* the conversation in sign language AI.

## 1. Introduction

Applications of machine learning (ML) and artificial intelligence (AI) to sign languages have exploded over the past few years. As large-scale sign language datasets emerge, a growing number of works apply data-driven AI methods from computer vision and natural language processing to solve various problems including sign language recognition, translation, and generation (Bragg et al., 2021; Yin et al., 2021; Börstell, 2023).

At the same time, the field has been shaped by systemic barriers causing the historical and present exclusion of Deaf<sup>1</sup> people from it (Angelini et al., *in press*). This includes the ableism and audism that shapes perceptions of Deaf communities and signed languages, as well as larger trends in STEM education that exclude Deaf individuals from being involved in research about them. Börstell (2023) shows that as many as 12% of papers in sign language computing contain basic ableist terms, double the incidence of such terms linguistics papers.

Towards more equitable research, previous work has identified major issues in papers, and issued recommendations on how to improve sign language AI research from multiple perspectives, including ethical considerations in datasets, linguistic as-

pects, and community engagement (e.g., Fox et al. (2023); De Sisto et al. (2022); Bragg et al. (2021); De Meulder (2021)). While these efforts are critical to addressing the ableism and audism that permeates the field, they generally focus on individual interventions encouraging authors to do better.

In our work, we reasoned that the systemic impact of excluding Deaf researchers from sign language AI research may be more subtle, and that a critical interrogation is needed of the assumptions and rhetoric that shape the research questions and methods in the field. In principle, even if each individual paper and research project followed best practices in responsible (sign language) AI, the collective direction of the field may still be misaligned with the interests and perspectives of most Deaf stakeholders. Collectively, what problems and aspects of signed languages are considered worth studying, and who decides such?

In other emerging fields, critical literature reviews have been crucial in redirecting research (e.g., Mack et al. (2021); Spiel et al. (2022); Froehlich et al. (2010)). Inspired by these works, we conducted a hybrid literature review and position paper analyzing over 100 papers in sign language AI.

Our analysis identifies systemic biases in the current state of sign language AI research. We show that the majority of papers are motivated by solving perceived communication barriers for Deaf individuals, use datasets that do not fully represent Deaf users, lack linguistic grounding, and build upon flawed models. From these results, we take the position that the field suffers from a lack of intentional inclusion of Deaf stakeholders. Lacking meaningful and ongoing input from Deaf stakeholders, the field

---

<sup>1</sup>We use 'deaf' to refer to audiological status, and 'Deaf' to refer to cultural identities. While the field of Deaf Studies is moving away from the use of deaf vs. Deaf (Kusters et al., 2017a), here we prefer a more explicit signposting of identity. While we aim to be precise, the miscible nature of identity means at times, our usage is interchangeable, but our intent is not to use terms as a means to exclude.



is instead driven by what approaches and modeling decisions are the most convenient. We end with a call to action: the field must make space for Deaf researchers to *lead* the conversation in sign language AI.

## 2. Positionalities and lived experiences

Our analysis and positions are shaped strongly by our identities and positionalities. We are a group of five researchers: we all identify as deaf, Deaf or hard-of-hearing (DHH). Two of us are white, three are Asian. Our interdisciplinary team spans a range of fields and research interests, including machine learning and computer vision, Deaf Studies and applied language studies, linguistics language documentation/corpora, phonetics/phonology, HCI and accessibility, psycholinguistics, language acquisition, developmental psychology, and cognitive science. We recognize that we come from positions of literacy and educational privilege, which may not be representative of Deaf communities. Our daily communication encompasses a blend of signed, written, and for some of us, spoken languages. Collectively, our linguistic repertoires include ASL, International Sign, NGT, VGT, KSL, English, Dutch, Gujarati and Hindi, along with other languages. Our experiences with assistive hearing technologies vary, with some of us having used hearing aids in the past while others continue to use them. We have varied lived experiences, but share the experience of growing up deaf or hard-of-hearing and going to mainstreamed schools for all or most of our education. Some of us grew up signing. For some of us, signing has been a part of our lives from an early age, while others began signing in their teenage years.

That all authors of this paper are DHH is intentional. Our aim from the outset was to approach this research from explicitly DHH positionalities and to bring different viewpoints. Since deaf people are the primary stakeholders in sign language technologies, we believed it essential to foster a space where DHH researchers could engage in open discussions about biases in ML applications to sign languages. The act of suggesting that hearing collaborators may be contributing to systemic bias seen in the field puts undue burden on DHH authors to carefully manage what they say. Because every member was DHH, we were able to openly discuss systemic bias and extend our discussion to not only include very clear instances of ableist works but also delve into the more subtle effects of ingrained biases in sign language AI research. Similar spaces created by other DHH scholars have generated insightful discussions of issues central to Deaf stakeholders (Kusters et al., 2017a; Chua et al., 2022; O'Brien et al., 2023).

## 3. Methods

### 3.1. Corpus creation

Sign language computation research lies at the intersection of Natural Language Processing, Computer Vision, and Human-Computer Interaction/Accessibility. As no dedicated venues centralize the majority of relevant work, we turned to arXiv, where computational researchers often share preprints of their work. We retrieved all papers containing the term “sign language” in CS field on arXiv, scoping our search to papers January 2021 to November 2023. This yielded 222 papers.

As our review focuses on sign language AI, we exclude works that exclusively study human factors. For papers in sign language AI, we focus on “receptive” sign language models, models that accept a recording or representation of sign language as input. Although work that focuses on sign language generation or avatars is also interesting and contributes to language understanding, these methods are relatively less developed (Yin et al., 2021). We reasoned focusing on receptive models would provide more diversified design decisions for analysis while reducing the volume of papers. We also exclude works that do not center sign language (e.g., uses sign language to demonstrate how methods generalize). Since our work focuses on sign *language*, we include work that focuses on fingerspelling only if they explore fingerspelling in the context of a longer sentence, or if the work (erroneously) claims fingerspelling to be a complete language system. We exclude reviews, theses, and non-English works.

Three authors reviewed paper abstracts against our inclusion criteria. Initially, two authors were assigned to each abstract. If there was a disagreement, a third author broke the tie. After filtering through inclusion criteria, we had 137 papers.

A limitation of arXiv is that works have not necessarily been peer reviewed. We only include published works from 2021-2022 (excluding 26 papers). As 2023 arXiv papers might be currently undergoing review, we include all preprints from that year that match our inclusion criteria. This gave us a total of 111 works for our systematic analysis.

### 3.2. Systematic Literature Review

We developed a codebook iteratively through discussion between authors (see appendix). We track the datasets used in each paper alongside inputs to models and outputs of the model (i.e., labels). We also note any prior models that papers build on (i.e., pretraining). We additionally read the abstracts and introduction to understand how the paper is motivated. Two annotators coded each paper, and disagreements were resolved by a third annotator.



## 4. Results and Discussion

We excluded 10 papers from our initially compiled list on further review as we found they did not match inclusion criteria. Our review thus consisted of a total of 101 papers, 21 from 2021 (peer-reviewed), 29 from 2022 (peer-reviewed), 51 from 2023 (arXiv). Most of these works focused solely on sign language recognition or translation as their main task, with a few looking at additional tasks like segmentation, sign spotting, etc.

Of the 101 papers in our review, we find that 60 work with continuous sign language datasets, 26 work with isolated sign language datasets, 3 with a combination of isolated and continuous sign language datasets, and 11 work with fingerspelling data. Most datasets used are publicly available. Seven works collect their own private dataset. Below we discuss themes from our systematic review.

### 4.1. Papers are motivated by perceived communication barriers

In our review, we find that 64 papers primarily motivate their work as addressing barriers in communication between deaf people and hearing society or spoken language resources. Navigating a hearing world and resulting communication barriers are undeniably a central component of the lived deaf experience. However, sign languages are not merely “communication tools” (Hu et al., 2023b), they are full languages, with a long history of being recognized as such (De Meulder et al., 2019). When ML research focuses singularly on the role sign languages play in provisioning access, it overlooks the history and diverse lived experiences of Deaf people, and misses out on exciting avenues for research, as we discuss below.

We find in most papers, the description of communication barriers encountered by deaf individuals either implies or directly establishes an inherent connection between sign language use and hearing ability. First, many papers claim that sign languages are the “primary form” (Walsh et al., 2023) or “natural means” (Varol et al., 2021) of communication for deaf people. However, not all deaf individuals know and use a signed language, and the signing communities extends beyond those who identify as Deaf. Even as individuals should have the right to self-determine what communication modalities they use in what contexts, the systemic suppression of sign languages means that many deaf people are not given sign language as an option in the first place. By presenting an oversimplified claim that “deaf people use sign language”, authors fail to pay credence to this long-standing oppression (as well as movements seeking equal status for signed languages) that complicate this relationship (Murray et al., 2019).

Second, there is a frequent narrative in the papers that suggests the primary hurdle in communication between ‘deaf’ and ‘hearing’ people is the ‘lack of a shared language’, with some papers claiming that deaf people largely lack fluency in written languages (e.g., “the globe’s [430 million DHH people] largely do not benefit from modern language technologies” (Wang and Nalisnick, 2023)). This framing diminishes the multilingual and multimodal capabilities of deaf people (Kusters et al., 2017b). Often, deaf and hearing people *do* share a common language, but deaf people might not have physical access to auditory languages. Most sign languages do not have a commonly used written form and so deaf signers often learn to read and write in another language (Gärdenfors, 2021), even as some face (and overcome) barriers in acquisition of spoken languages. Additionally, by fixating on how deaf people communicate exclusively, this framing portrays communication as one-sided when it is usually reciprocal and multimodal. ‘Communication’ for deaf people is much more complex than a mere translation between signed and spoken languages.

Third, perceived communication barriers are often used to argue that deaf people are not included into hearing society, and therefore experience adverse consequences. For example, in their discussion of broader impact, Hu et al. (2023a) state that deaf people may “feel isolated, lonely, or [have] other mental health issues when they face the communication barrier in daily life”. While it is true that inaccessibility impacts deaf people on a systemic and individual level, claims like these portray deaf people as deficient and in need of technological interventions (termed by Morozov (2013) as ‘technosolutionism’), instead of more accurately recognizing that most deaf individuals already have developed strategies to navigate hearing society, and that any emerging technology will at least initially only be a small supplement to these strategies. Thus, this framing of deaf individuals is ideological, allowing authors to overstate the importance of their contributions to the daily lives of deaf people, at the expense of diminishing their existing repertoires.

We note that not every paper that focuses on communication barriers frames poorly. For example, Hossain et al. (2023) are careful to scope their claims to barriers in STEM education and design a method well aligned with the application. However, we believe there are two distinct issues: first, in our reading, the majority of papers that do motivate their work as addressing communication barriers do have oversimplified or inaccurate views. But second is the overall proportion of papers in the field that focus on mitigating communication barriers.

Addressing the second issue, in our view, this means the field disproportionately focuses on a single story: mitigating accessibility barriers, which is

primarily understood to be “deaf people’s access to spoken language”. This means that receptive sign language models are mostly studied in the context of translation, overemphasizing the role of spoken language. While this is an important issue, it is not the only framework in which sign language recognition can occur. In our review, we find a few works that are motivated by exploration of sign language as a language in its own right, including models that annotate phonology (Tavella et al., 2022), or predict the iconicity of signs (Hossain et al., 2023), but these are far less represented than translation works. Sign languages *are* different in many ways than spoken languages, and rather than considering these differences as inherent limitations that make building sign language technologies difficult, there is an opportunity to develop AI technologies that understand and center these differences to further our scientific understanding of the human capacity of language. For example, as we further discuss in Section 4.3, most translation annotation schemes focus on flattening phonological differences between users to prioritize semantics, but differences in phonology can induce differences in meaning, as well as connect to the identity of the signer. Applications like these are currently underserved by sign language AI.

#### 4.2. Models use datasets misaligned with target users

Across all papers, we identified 43 different publicly available sign language datasets. 16 datasets use solely DHH contributors, 3 datasets use solely interpreters, 11 datasets include a mix of contributors, and 12 datasets do not specify contributor qualifications. While this heterogeneity in dataset contributors seems promising at the surface, it raises several concerns. First, most papers claim to build technologies to solve communication barriers for deaf people, but many (12 of 43) datasets do not disclose who they collect data from. This indicates an underlying assumption: that everyone signs the same way or that variations in signing are insignificant. We unpack additional concerns below.

Second, even as datasets are diversified in terms of contributors, their usage is not. The three datasets that use interpreters only (Albanie et al., 2021; Forster et al., 2014) are long-standing benchmarks in the field, and are used by 41 of the 60 continuous sign language recognition works in our systematic review. All three of these datasets are continuous sign language and draw from existing media broadcasts. While these works offer large-scale annotated datasets to advance sign language recognition (which has known to be constrained by lack of data), the question arises whether it is appropriate to use interpreted datasets as source ma-

terial to develop sign language AI. First, the majority of sign language interpreters are hearing users who may not sign in a manner that aligns with usage patterns in Deaf communities. Instances have been documented where Deaf viewers face challenges in understanding the interpreters in the same broadcasts used for ML purposes (Alexander and Rijckaert, 2022). Secondly, the nature of scripted and interpreted language use, especially under the constraints of simultaneous interpreting, diverges significantly from language in the wild. This may result in a distorted representation of sign languages in AI systems (see also SignOn (2022)). We note that authors of some of these datasets discuss limitations – e.g., Albanie et al. (2021) (BOBSL dataset) remark on “translationese” extensively – but most works that use these datasets do not. These distortions have broader implications. Deaf end users may find themselves compelled to adjust their sign language use to accommodate the limitations of AI technologies trained on this data, a form of linguistic subordination to technology.

More recent datasets have recognized this gap between training data and target users, and sought to collect more representative data – ASL Citizen (Desai et al., 2024) and Sem-Lex (Kezar et al., 2023) are both large scale isolated sign language recognition datasets of ASL, and aim to collect data from “fluent” DHH signers. While this is an improvement, details in how participants were recruited reveal that the notion of “fluency” is more subjective than what is discussed in either paper. ASL Citizen claims to recruit “fluent signers” from “trusted groups” but does not state what/who these are. In contrast, Sem-Lex defines “fluent” signers as those who acquired sign language in childhood. While people who acquired sign language in childhood are a portion of contemporary Deaf communities, it is not the only group, and not even the largest one. 95% of deaf children are born to hearing families. Often these children do not learn sign language until later in life or at all, because medical practitioners often discourage parents from using a signed language (Murray et al., 2019). This illustrates the ideological meaning of “fluency”. While later or different acquisition paths means they might sign differently from the ideological “norm”, excluding them from datasets means we exclude them as users of designed technologies. While targeting subsets of the community can help scope data collection, our concern is how this bias is framed: Sem-Lex argues for data representative of “deaf signers” in general, without explicitly discussing how their data may not be representative of many signing Deaf people. Without this disclosure, we worry this may lead to applications that inadvertently marginalize a large proportion of Deaf communities.

Overall, perhaps the biggest driver in mis-

matches between data and applications is the opposing goals of data as needed for machine learning applications and language as it happens in the world. First, finding an optima for machine learning necessitates scoping multi-dimensional and nuanced realities to something neat and tractable. Datasets make decisions about what variation is desirable to collect, and what is out-of-scope for a particular dataset. For example, ASL Citizen considers variation in background, illumination, and camera angle of recorded videos desirable, and Sem-Lex considers signer diversity across race and gender axes. At the same time, the prompting and labelling procedures in both datasets both seek to minimize label noise for signs for each category. In ASL Citizen, contributors are prompted to copy a seed signer’s production of a sign, instead of providing their own sign for a concept. Similarly, in Sem-Lex, if a contributor provides a sign that is not included in a pre-defined corpus, it is discarded. This creates tension in the decision to collect a racially diverse dataset: even if Deaf people of color are represented, if a dataset only retains signs they produce that are present in dictionaries historically biased towards language used by white people (Hill, 2023), signs they use within their own communities may be discarded.

Clean data and high quality annotations are therefore in direct tension with procedures that foster agency and authenticity from signing contributors. This tension plays out in many different ML fields (Bender and Friedman, 2018), but we are more concerned with how characterizations for desirable and excluded variation for datasets tie to a larger societal rhetoric of “good” and “bad” language. Revisiting our earlier discussion of fluency as an ideal, we note the concept of fluency is frequently entangled with notions of racial and ableist privilege, often being contingent upon closeness to whiteness and normative physical ability (Henner and Robinson, 2023). Without a critical examination of what constitutes “fluency”, there is a risk of elevating those who, by virtue of early exposure to sign language and alignment with privileged identities (e.g., racial, able-bodied), are considered the “purest” or most “ideal” users (also see ‘native’ signer bias discussed in Hochgesang et al. (2023)). This paradigm risks overshadowing the diverse linguistic realities of deaf people and can again perpetuate a form of linguistic subordination to technology, where users are compelled to conform their signing to that of the “ideal”, “fluent” model. This further overlooks the varied experiences of Deaf people with additional disabilities that might influence their interaction with sign language AI technologies, or even for Deaf people considered “fluent” if they need to modify their signing (e.g., they’re signing one-handed because they’re holding an object), in

contravention with a goal of accessible design.

But second, the need for large scale training data may engender reliance on more scalable data collection procedures (Bender et al., 2021) (e.g., collecting data from hearing interpreters, scraping from publicly available videos on the Internet, using subtitles) and result in suboptimal datasets that do not capture language as used by deaf people. We discuss this more in the next section, but for now, we ask the question: who gets to decide whether using or collecting more data outweighs the possibility that data may lead to biases that marginalize (Bender et al., 2021)?

### 4.3. Labels lack linguistic foundation

Next, we looked at the annotation schemes used by models, which we found to be a good proxy for understanding how models use (or misuse) prior linguistic knowledge. We find that half of the papers (51) rely on glosses – a written language representation of signed language intended to preserve original meaning and structure (Comrie et al., 2008) – as either their main output or intermediary representation. Specifically, we find 30 papers that use glosses alone, with an additional 17 using glosses alongside spoken language translations, 4 using glosses alongside phonological features or other annotations.

We find that sign language AI research has adopted the use of glosses without discernment, and without following best practices pioneered in linguistics (Hodge and Crasborn, 2022). Glossing conventions in linguistics are closely tied to projects: there is no singular gloss system, and gloss systems vary depending on the theoretical framework and questions of the research team. This similarly happens in sign language datasets, regardless of whether the gloss system is intentionally designed, or a consequence of data processing. For example, WLASL (Li et al., 2020) (an ISLR dataset) merges gloss systems from different scraped online resources, and this leads to a final gloss system largely based on their English literal - in this gloss system, the sign for **PRESENT** meaning gift, and **PRESENT** meaning time are represented by the same gloss<sup>2</sup>. This is distinguished from ASL Citizen and Sem-Lex, which use a gloss system from ASL-LEX (Sehyr et al., 2021), which distinguishes signs by their semantics (e.g., **BOW\_1** meaning hair ornament, and **BOW\_2** meaning archery are given distinct glosses<sup>3</sup>). There are still other glossing systems that would be useful for and employed

<sup>2</sup>PRESENT - gift - [handspeak.com/word/3783/](https://handspeak.com/word/3783/)  
PRESENT - time - [handspeak.com/word/2751/](https://handspeak.com/word/2751/)

<sup>3</sup>BOW\_1 - [asl-lex.org/visualization/?sign=bow\\_1](https://asl-lex.org/visualization/?sign=bow_1) BOW\_2 - [asl-lex.org/visualization/?sign=bow\\_2](https://asl-lex.org/visualization/?sign=bow_2)



by linguists in some contexts (e.g., those that make finer distinctions between phonological variants of the same sign), that we did not find represented in current sign language AI research. Critically, glosses cannot represent all linguistic phenomena in signing, e.g., signs that point or depict, name signs, etc. Researchers often rely on internal or current practices for additional conventions.

Second, while glosses generally make source languages accessible to those in the field who may not be fluent in both languages, they do not stand alone as a complete representation, and lose meaning like any translation. In linguistic research, glosses often accompany the source language as to provide some access to meaning for those not fluent. Unfortunately, in sign language research, glosses are often used as the *only* representation of signs, without any direct link to the source (be it video, photos or drawings), even when the issues with this representation are known – a phenomenon called the “tyranny of glossing” (Hochgesang, 2019, 2022b)<sup>4</sup>.

Here, we are concerned that the use of glosses in sign language AI research goes one step further, where many papers treat glosses as an actual translation, rather than a context-dependent representation. This is evidenced by several observations. First, virtually no paper describes the underlying design of the gloss system they are predicting. Without knowing what is being predicted, models lack usefulness for linguistic applications. Second, many papers build predictors on several independent datasets. We consider this to be predicting several independent, if correlated and not fully disclosed, tasks - e.g., WLASL predicts the English word associated with a sign, whereas ASL Citizen predicts semantic categories of phonologically distinct signs. However, many of these papers claim these predictors are accomplishing sign language translation, effectively claiming these distinct and disparate gloss systems as complete representations of sign language. Third, for continuous sign language, the field often approaches sign language translation as a two-phase pipeline consisting of movement from sign2gloss and gloss2text. However, discussion is often not given to how the gloss system may bottleneck information (e.g., if spatial and temporal components are represented).

Even works that do not use glossing may face the same issues. 11 papers do not specify what kind of annotation system they use, but attempt ISLR through a classification framework. The target here impacts task difficulty and the final application. We also find papers that use different systems – 4 works use phonological features, and 5 use other

notation systems like HamNoSys – systems which are also specialized, noisy, and tied to specific theoretical perspectives on signs (Hochgesang, 2014). Our point is not that glosses are inherently bad, rather that they are partial and subjective representations of sign language and deeply shape the task at hand. When researchers focus on improving model performance without contextualizing what they are even predicting, they fail to engage with a core part of the research. ML scholars need to be explicit with their design choices and articulate trade-offs between systems.

We also note a growing trend of end-to-end translation, where works use spoken language translations as targets (18 works in our review). This is largely motivated by the difficulty and expense in acquiring high quality annotations for sign language data. These works instead often rely on subtitles for supervision. While one might think this avoids the issues above, it adds other considerations. First, there is no guarantee that the subtitles reflect the same content or order of content, for a number of reasons. In simultaneous work, the captionist or interpreter may miss content; in translation, the interpreter may need to inject additional context depending on audience; and if captions are automated, biases from technology can be injected (e.g., automated captioning struggles with technical terms and accents). But even in situations where the subtitles reflect reliable translations, translation itself may not be perfect. For example, the lyrics of a song used in a sign language music video are technically accurate, but will miss the expressive art of the signer. Generalizing on this example, by relying heavily on spoken corpora, we limit ourselves only representations that align with spoken language conventions, paralleling issues raised in Section 4.1. Finally, that most work focuses on mapping sign languages to spoken languages (including glossing) is uncomfortable, because it echoes misconceptions that sign languages are not independent, but analogues of spoken languages. As mentioned in section 4.1, translation is not the only possible framework under which sign language recognition can occur, and there is opportunity to center other tasks like sign language understanding instead.

Overall, despite sign language modeling being framed as an computer vision *and* natural language processing problem, we find there is a lack of linguistic awareness and incorporation of linguistic knowledge into research approaches. This leads to researchers appropriating annotation schemes without context (such as glossing), prioritizing ease rather than quality (such as subtitles), and over-relying on semantic representations (tied to spoken languages, rather than other representations that offer other applications).

---

<sup>4</sup>(with gratitude to Börstell for coining “Glossgesang”) [twitter.com/c\\_borstell/status/117749859992610823?s=20](https://twitter.com/c_borstell/status/117749859992610823?s=20)



#### 4.4. Modeling decisions inherit biases

Next, we looked at machine learning modeling decisions. Of the 101 papers in our review, we found that 59 models use vision-based inputs (i.e., RGB video or images), 34 use pose-based inputs (i.e., joint keypoints estimated from videos by a pose extractor), and 10 use other input representations (e.g., manually assigned features or 3D sensor data). Note some works use multiple inputs.

Data-driven AI-approaches typically rely on large amounts of annotated data to train. As most sign language datasets are small, many works will employ transfer learning approaches, where sign language models will fine-tune or rely on outputs from previous models pretrained in another setting, where data is more abundant. However, transfer learning is not without its risks: pretraining can introduce biases into models that are inherited by fine-tuned models (Wang and Russakovsky, 2023).

From this perspective, it is concerning that 34 of the papers use pose-based inputs, which are extracted from pre-trained pose estimators (Lugaresi et al., 2019; Cao et al., 2017; Fang et al., 2022). These are models not trained on sign language data, but using action or gesture videos. Moryossef et al. (2021) show failure models and biases when applying them to sign language: for example, hand-shapes in sign language are typically much more fine-grained than what these models encounter in pre-training. Furthermore, by construction, many pose-based models exclude information necessary to understand sign language: for example, even though MediaPipe (Lugaresi et al., 2019) extracts facial landmarks, Selvaraj et al. (2021) advocate for the use of a reduced set of keypoints that include no information about facial expression, even in continuous sign language settings where the face is critical to grammar.

Similarly, many of the vision-based models (42 of 59 models) also employ pre-training. 24 of these models only pre-train on non-sign language datasets (with ImageNet (Deng et al., 2009), a natural image dataset, and Kinetics (Carreira and Zisserman, 2017), a human action dataset being most common). Again, it is unclear what biases are inherited with this approach: previous work by Desai et al. (2024) shows that models pre-trained on Kinetics provide no capability to recognize isolated signs beyond random chance, and work by Shi et al. (2022) suggests that pre-training on ImageNet may in some cases, degrade performance. While the other 18 models do explore pre-training on sign language datasets instead, the majority of these works pre-train on BSL (8 models) or ASL (8 models). These models often then evaluate on other sign languages, and although we consider this pre-training to be a closer domain than e.g., action videos, it is unclear if this introduces any biases

in phonology shared between the sign languages versus distinct. In our analysis, we identified no paper that provided a quantitative analysis of potential biases from pre-training: even though as papers compare pre-training versus training from scratch (Jang et al., 2022) or different pre-training datasets (Shi et al., 2022), all papers report overall metrics on datasets exclusively, without seeking to understand if performance increases come with trade-offs (e.g., reporting metrics class-by-class to understand if improving recognition of some signs comes at the expense of others).

Beyond pre-training, a second sub-theme that we observed is that even as some papers claim to produce general methods, it is unclear if methods are correcting issues cascading from previous design decisions. An interesting case example is in Zuo et al. (2023), which argues that semantic similarity in English glosses can be used to improve sign language recognition, as sometimes signs related in meaning share phonology. However, to demonstrate this claim, this paper relies primarily on internet-scraped datasets that rely upon English glosses to merge and distinguish signs, including WLASL (Li et al., 2020). Our exploration of this dataset suggests that this procedure creates artifacts where distinct glosses refer to identical signs in ASL (e.g., “DORM” and “DORMITORY”), and it is unclear if improvement from the proposed method is due to correcting these artifacts versus general linguistic properties of sign languages.

Overall, we observe the majority of sign language AI works build off previous methods, with known issues and flaws in how they represent sign language. While this point is understandable because re-inventing every design aspect of a new sign language model is unreasonable for any individual paper, this means that issues are inherited by future models, often uncritically. Echoing perspectives from previous sections, we argue that in many cases this is because authors lack the linguistic expertise to fully identify where modeling decisions not be representative or general. This creates systemic biases in modeling that align with decisions made due to convenience (e.g., it’s easier to use existing pose-based models, rather than training one specialized to sign language), but ultimately become standards as new papers do not re-assess if these design decisions align with sign language, but uncritically adopt them as defaults. While foregoing pretraining entirely might not be feasible given the data gap, works can analyse impacts of pretraining more closely and explore how one might mitigate inherited biases.

## 5. Calls to Action

Synthesizing our results, we take the position that as a field, sign language AI research lacks *intentionality*: collectively, problem formulation and model design is not guided by what best aligns with Deaf stakeholder interests or growing trends in sign language research that center the complexities of lived deaf experiences. In the absence of these guiding principles, these decisions are left to researcher preference and ease. We showed that in spite of a range of possible problem formulations, datasets, targets, and models, most works narrow to a few defaults. Although our point is that this is problematic even if every paper is well-executed, we expose numerous issues to demonstrate that these biases are likely induced by positionality, as most research is led and conducted by hearing non-signing researchers. That most research is motivated by communication barriers is tied to the issue that many researchers view deaf people as being ‘deficient’. That most papers use datasets or prediction targets that misalign with broader Deaf languaging patterns connects to how many authors lack linguistic knowledge and actual engagement with Deaf communities. Some of these misaligned decisions are now baked-in as standards, such as the use of interpreter-only datasets as benchmarks, or the use of pretrained models without fully understanding their biases. These misalignments have the potential to marginalize the very target users of sign language technologies. Moreover, as Deaf signing communities are a wide spectrum, they may marginalize subsets of the community even as they serve others.

Towards addressing this systemic issue, we advocate that the field foster Deaf leadership. Previous works have advocated for including Deaf collaborators (Yin et al., 2021), and while we agree that Deaf-hearing collaboration is essential to make meaningful progress in the field, we also believe that including Deaf people in each individual project is not a structural solution. First, just including Deaf collaborators does not necessarily mean they are driving the research agenda. In most cases, they are not. In the first-hand experience of the authors of this paper, Deaf researchers are often only asked to collaborate often well after the idea has been conceived, the team built, the research conducted, or even near the project write-up as sometimes the sole “deaf” person. In this paper, we showed that there are often tensions between how to allocate limited resources in projects and making decisions that are linguistically and culturally appropriate. Currently, most of these decisions are made by hearing (often non-signing) researchers, and sometimes this is done even without awareness that an impactful decision is being made. “Lead-

ership by the most impacted” is one of the core principles of Disability Justice (Berne et al., 2018) : even if Deaf researchers may not have all the answers in these complex trade-offs, enabling us to lead research means these decisions are at least being made by those with a larger stake.

But second, Deaf researchers are underrepresented in the field, and even if exclusionary structures are fully addressed, may still persist as a minority for demographic reasons. Asking DHH scholars to be involved in each individual project creates burden given the overwhelming number of sign language AI works relative to the number of DHH researchers, and may distract them from other priorities or create tensions where they feel declining a project harms their community (Angelini et al., in press). Instead, the field needs to contend with how to amplify Deaf perspectives, even as they may continue to form a minority of research outputs. Towards this end, hearing researchers should reassess their role in work involving Deaf signing communities. Rather than being the ones to dictate the agenda and be the public face, hearing researchers can transition these opportunities to Deaf researchers, and instead switch to a role of supporting Deaf researchers like taking on the responsibility of accessibility or promoting their training.

For this to be possible, all researchers in sign language AI research need to be transparent about their positionalities. This imperative extends beyond a ‘confession before a crime’, aspiring instead to weave positionality deeply into the research, enhancing transparency and underscoring the impact of researchers’ backgrounds, experiences, privileges, and biases in their work. Transparency about one’s positionalities is an increasingly recognized practice in sign language linguistics, sociolinguistics, interpreting studies, and Deaf Studies research (Hou, 2017; Kusters et al., 2017a; Kusters and Lucas, 2022; Mellinger, 2020; Hochgesang, 2022a), where the lived experience of researchers (DHH or hearing) can significantly differ from those of their participants in aspects such as ethnicity, race, other disabilities, and educational and linguistic backgrounds. Positionality statements are not a standard for sign language AI research (although some works informally disclose (Bragg et al., 2021; Desai et al., 2024)), but given how the field contends with similar issues with potential mismatches between researchers and target users, we recommend it become adopted as practice.

At the same time, we are cautious about our call for Deaf leadership. While we believe it is a meaningful step forward, it is not a full solution in itself, and followed uncritically, it risks corruption of the very principles we issue this recommendation under. We’ve noted that calls for and projects

that claim Deaf collaboration or leadership have become tokenizing (De Meulder and Kusters, 2021). We worry that our call for Deaf leadership may be similarly impacted. Without carefully considering whose voices to include, how to meaningfully build consensus, and how to reconcile disagreements, attention might focus on those who already have the most power, glossing over inequalities within the community. Deaf researchers themselves must acknowledge there are gaps, and Deaf leadership must come from a wide range of perspectives and backgrounds. We are careful to note our own positionalities (e.g., educational and literacy privilege). We further found critiques of our own work upon reflection (e.g., ASL Citizen, which two authors on this paper worked on). Just because we are DHH doesn't mean we are immune to participating in systemic biases.

Thus, our call for Deaf leadership is intended to be a call for ongoing conversation, one in which we continuously re-evaluate how positionality influences research, and where stakeholders need to be in charge of decisions. For example, even as we ask hearing researchers to transfer visibility and accountability to Deaf researchers, to what extent does this depend on the project, the discipline(s), and other people involved? And to which Deaf researchers? Even now, these are questions we do not fully have the answers to. But to find answers, there first has to be a conversation taking place, which is currently absent from large swaths of the field. We invite all sign language AI researchers to join the conversation.

## 6. Acknowledgements

This work was supported by Center for Research and Education on Accessible Technology and Experiences (CREATE) and partially funded by a gift from Microsoft. We would like to thank Cassandra Kim for valuable help with annotation in our systematic analysis. We would also like to thank Danielle Bragg, Hal Daumé III, and Mary Gray for continued discussions that sparked ideas that informed this work.

## 7. Bibliographical References

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Dhoest Alexander and Jorn Rijckaert. 2022. [News 'with' or 'in' sign language? case study on](#)

[the comprehensibility of sign language in news broadcasts](#). *Perspectives*, 30(4):627–642.

Robin Angelini, Katta Spiel, and Maartje De Meulder. in press. Bridging the gap: Understanding the intersection of deaf and technical perspectives on signing avatars. *A. Way, D. Shterionov, C. Rathmann & L. Leeson (eds.), Sign Language Machine Translation*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Patricia Berne, Aurora Levins Morales, David Langstaff, and Sins Invalid. 2018. Ten principles of disability justice. *WSQ: Women's Studies Quarterly*, 46(1):227–230.

Carl Börstell. 2023. Ableist language teching over sign language research. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 1–10.

Danielle Bragg, Naomi Caselli, Julie A Hochgesang, Matt Huenerfauth, Leah Katz-Hernandez, Oscar Koller, Raja Kushalnagar, Christian Vogler, and Richard E Ladner. 2021. The fate landscape of sign language ai datasets: An interdisciplinary perspective. *ACM Transactions on Accessible Computing (TACCESS)*, 14(2):1–45.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

M Chua, Maartje De Meulder, Leah Geer, Jonathan Henner, Lynn Hou, Okan Kubus, Dai O'Brien, and Octavian Robinson. 2022. 1001 small victories: Deaf academics and imposter syndrome. In *The Palgrave handbook of imposter syndrome in higher education*, pages 481–496. Springer.



- Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2008. Leipzig glossing rules. conventions for interlinear morpheme-by-morpheme glosses. max planck institute for evolutionary anthropology, leipzig.
- Maartje De Meulder. 2021. [Is “good enough” good enough? ethical and responsible development of sign language technologies](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 12–22, Virtual. Association for Machine Translation in the Americas.
- Maartje De Meulder and Annalies Kusters. 2021. [Twitter thread](#).
- Maartje De Meulder, Joseph J Murray, and Rachel L McKee. 2019. *The legal recognition of sign languages: Advocacy and outcomes around the world*. Multilingual Matters.
- Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, Goggi S, Isahara H, Maegaard B, Mariani J, Mazo H, Odijk J, Piperidis S, editors. LREC 2022, 13th International Conference on Language Resources and Evaluation; 2022 June 20-25; Marseille, France. Paris: European Language Resources; 2022. 10 p*. European Language Resources Association.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE.
- Aashaka Desai, Lauren Berger, Fyodor Minakov, Nessa Milano, Chinmay Singh, Kriston Pumphrey, Richard Ladner, Hal Daumé III, Alex X Lu, Naomi Caselli, et al. 2024. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *Advances in Neural Information Processing Systems*, 36.
- Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.
- Neil Fox, Bencie Woll, and Kearsy Cormier. 2023. Best practices for sign language technology research. *Universal Access in the Information Society*, pages 1–9.
- Jon Froehlich, Leah Findlater, and James Landay. 2010. The design of eco-feedback technology. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1999–2008.
- Moa Gärdenfors. 2021. [The writing process and the written product in bimodal bilingual deaf and hard of hearing children](#). *Languages*, 6(2).
- Jon Henner and Octavian Robinson. 2023. Crip linguistics goes to school. *Languages*, 8(1):48.
- Joseph C. Hill. 2023. [Overrepresentation of whiteness is in sign language as well: A commentary on “undoing competence: Coloniality, homogeneity, and the overrepresentation of whiteness in applied linguistics”](#). *Language Learning*, 73(S2):312–316.
- Julie Hochgesang. 2022a. [Documenting signed language use while considering our spaces as a deaf\\* linguist](#).
- Julie Hochgesang. 2022b. Managing sign language acquisition video data: A personal journey in the organization and representation of signed data. *The Open Handbook of Linguistic Data Management*, pages 367–383.
- Julie A Hochgesang. 2014. Using design principles to consider representation of the hand in some notation systems. *Sign Language Studies*, 14(4):488–542.
- Julie A Hochgesang. 2019. Tyranny of glossing revisited: reconsidering representational practices of signed languages via best practices of data citation. In *TISLR13, the 13th Conference of Theoretical Issues in Sign Language Research, Hamburg, Germany (September 26–28, 2019)*.
- Julie A Hochgesang, Ryan Lepic, and Emily Shaw. 2023. [W \(h\)ither the asl corpus?: Considering trends in signed corpus development](#). In *Advances in Sign Language Corpus Linguistics*, pages 287–308. John Benjamins.
- Gabrielle Hodge and Onno Crasborn. 2022. Good practices in annotation. In *Signed language corpora*, pages 46–89. Gallaudet University Press.
- Sameena Hossain, Payal Kamboj, Aranyak Maity, Tamiko Azuma, Ayan Banerjee, and Sandeep Gupta. 2023. Edgcon: Auto-assigner of iconicity ratings grounded by lexical properties to aid in generation of technical gestures. In *Proceedings*



- of the 38th ACM/SIGAPP Symposium on Applied Computing, pages 3–10.
- Lynn Hou. 2017. Negotiating language practices and language ideologies in fieldwork: A reflexive meta-documentation. *Innovations in deaf studies: The role of deaf scholars*, pages 339–360.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023a. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023b. Self-emphasizing network for continuous sign language recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 854–862.
- Youngjoon Jang, Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. 2022. Signing outside the studio: Benchmarking background robustness for continuous sign language recognition. *arXiv preprint arXiv:2211.00448*.
- Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. The semlex benchmark: Modeling asl signs and their phonemes. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–10.
- Annelies Kusters, Maartje De Meulder, Dai O’Brien, et al. 2017a. Innovations in deaf studies: Critically mapping the field. *Innovations in deaf studies: The role of deaf scholars*, pages 1–53.
- Annelies Kusters and Ceil Lucas. 2022. Emergence and evolutions: Introducing sign language sociolinguistics. *Journal of Sociolinguistics*, 26(1):84–98.
- Annelies Kusters, Massimiliano Spotti, Ruth Swanwick, and Elina Tapio. 2017b. Beyond languages, beyond modalities: Transforming the study of semiotic repertoires. *International Journal of Multilingualism*, 14(3):219–232.
- Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.
- Kelly Mack, Emma McDonnell, Dhruv Jain, Lucy Lu Wang, Jon E. Froehlich, and Leah Findlater. 2021. What do we mean by “accessibility research”? a literature survey of accessibility papers in chi and assets from 1994 to 2019. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Christopher D Mellinger. 2020. Positionality in public service interpreting research. *FITISPos International Journal*, 7(1):92–109.
- E. Morozov. 2013. *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don’t Exist*. Penguin Books Limited.
- Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3434–3440.
- Joseph J Murray, Wyatte C Hall, and Kristin Snoddon. 2019. Education and health of children with hearing loss: the necessity of signed languages. *Bulletin of the World Health Organization*, 97(10):711.
- Dai O’Brien, Gabrielle Hodge, Sannah Gulamani, Kate Rowley, Robert Adam, Steve Emery, John Walker, et al. 2023. Deaf professionals’ perceptions of trust in relationships with signed/spoken language interpreters. *Translation & Interpreting*, 15(2):25–42.
- Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.
- Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2021. Openhands: Making sign language recognition accessible with pose-based pretrained models across languages. *arXiv preprint arXiv:2110.05877*.
- Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.12870*.
- SignOn. 2022. [Sign language technology: Do’s and don’ts](#).
- Katta Spiel, Eva Hornecker, Rua Mae Williams, and Judith Good. 2022. Adhd and technology

research—investigated by neurodivergent readers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Federico Tavella, Aphrodite Galata, and Angelo Cangelosi. 2022. Phonology recognition in american sign language. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8452–8456. IEEE.

Gul Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16857–16866.

Harry Walsh, Ozge Mercanoglu Sincan, Ben Saunders, and Richard Bowden. 2023. Gloss alignment using word embeddings. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.

Angelina Wang and Olga Russakovsky. 2023. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3957–3968.

Shuai Wang and Eric Nalisnick. 2023. Active learning for multilingual fingerspelling corpora. *arXiv preprint arXiv:2309.12443*.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14890–14900.

## A. Methods Supplementary and Datasets

We used an iterative process to develop questions for our systematic analysis, guided by an in-depth qualitative review of a few papers by all authors. Papers for this qualitative analysis were nominated by authors based upon individual authors’ beliefs that they were representative of current modeling work, or would generate multidisciplinary discussion. Our

final questions focused on four different themes: framing of the research in abstract or introduction, datasets used by the papers and other inputs for modeling, annotation or labeling schemes used for model outputs, and the use of pretrained models anywhere in the ML pipeline. Two annotators coded each paper, and a third annotator was called in to resolve disagreements. Two of the annotators have a background in ML and are familiar with reading such papers, one annotator has a background in psycholinguistics.

Of the 101 papers in our review, we find that 60 work with continuous sign language datasets, 26 work with isolated sign language datasets, 3 with a combination of isolated and continuous sign language datasets, and 11 work with fingerspelling data (8 focus on recognition from images, 3 study fingerspelling in a continuous signing context aka in-the-wild).

There are a total of 43 publicly available datasets used across our corpus (each used to varying degrees). Seven works collect their own private dataset. The sign languages studied in the public datasets include the following: American Sign Language (ASL), Deutsche Gebärdensprache (DGS), Chinese Sign Language (CSL), British Sign Language (BSL), Turkish Sign Language (TSL), Russian Sign Language (RSL), Indian Sign Language (ISL), Lengua de señas argentina (LSA), Greek Sign Language (GSL), Lengua de Signos Española (LSE), Arab Sign Language (ArSL), Bangla Sign Language (BdSL), Vlaamse Gebarentaal (VGT), along with some multilingual datasets (JWSign, SP-10). We note that along with disparities in who contributes data, not all sign languages are equally represented.

# Evaluating Inter-Annotator Agreement for Non-Manual Markers in Sign Languages

Lyke D. Esselink , Marloes Oomen , Floris Roelofsen 

University of Amsterdam  
Amsterdam, the Netherlands  
{l.d.esselink, m.oomen2, f.roelofsen}@uva.nl

## Abstract

This paper is part of a larger project that aims to create a standardized procedure for annotating non-manual markers (NMMs) in sign language data. The paper describes two approaches to evaluating inter-annotator agreement, the *event-based* approach and the *frame-based* approach, and uses a combination of these two approaches to evaluate the annotation guidelines introduced in Oomen et al. (2023). The evaluation reveals that for several labels in the annotation scheme inter-annotator agreement is rather low. This indicates that the annotations guidelines need to be further improved. We present concrete recommendations for how this may be achieved, and intend to implement these recommendations in future work. All data and analysis scripts are available.

**Keywords:** sign language, non-manual markers, annotation guidelines, inter-annotator agreement

## 1. Introduction

This paper is part of a larger project that aims to create a standardized procedure for annotating non-manual markers (NMMs) in sign language data. The initial steps we took as part of this project—developing annotation guidelines and creating a dataset annotated according to these guidelines by two annotators—were previously reported in Oomen et al. (2023). In the present paper, we report on the next step: a thorough evaluation of inter-annotator agreement, yielding substantial recommendations for improvement of the guidelines.

In Section 2, we outline our general motivations for developing a new protocol for annotating NMMs. Section 3 provides a brief summary of the first steps towards such a protocol as reported in Oomen et al. (2023). In Section 4, we describe two general methods for evaluating inter-annotator agreement which can be applied to sign language data. Section 5 discusses the results of applying these methods to our test dataset, leading to several recommendations for further improving our annotation guidelines. This is the main contribution of the paper. Section 6 discusses some methodological prospects and limitations of the evaluation methods we adopted, and Section 7 concludes. Before the bibliography, we provide pointers to all supplementary materials: the annotation guidelines, evaluation data, analysis scripts, and a technical report with extensive discussion of all results.

## 2. Motivation for the Larger Project

In sign languages, facial expressions, body movements, and other NMMs serve a wide range of linguistic functions, in addition to the gestural and

affective functions they may fulfil more generally.<sup>1</sup> There are plenty of examples in the literature tying particular NMMs (or clusters of NMMs) to particular grammatical functions (for a recent overview, see Wilbur, 2021). For instance, Bahan (1996) has argued that eye gaze (or head tilt in the case of first person) can be used to mark verb agreement in American Sign Language (ASL); Göksel and Keleşir (2013) have claimed that (forward or backward) head tilt in Turkish Sign Language marks interrogative mood while specific combinations of head tilt and head movement distinguish polar (forward + head nod) and content (backward + head-shake) questions; Wilbur and Patschke (1998) have proposed, again for ASL, that body leans are used to convey contrast at the prosodic, lexical, semantic, and pragmatic level. Works such as these provide highly valuable descriptive, analytical and theoretical insights, but they tend to be based on relatively small sets of examples, for which it is often unclear exactly how they were obtained or analyzed. The analyses also generally do not involve detailed qualitative annotation of NMMs, or the annotation procedure is not discussed.<sup>2</sup> Moreover, (individual) variation in NMMs use is often not considered. This means that many claims about NMMs and their properties and functions in sign languages still await robust empirical verification, which cannot be done without in-depth analysis of NMM patterns by means of careful annotation of linguistic data.

Facial expressions and other NMMs also play

<sup>1</sup>This section overlaps to a large extent with Section 2 from Oomen et al. (2023).

<sup>2</sup>Notable exceptions include Pendzich (2020) on lexical NMMs in German Sign Language and Lackner (2017) on the various functions of head and body movements in Austrian Sign Language.

an important role in multimodal communication, where they have been shown to be connected to a wide variety of semantic, pragmatic, and social functions (e.g., Bavelas and Chovil, 2018; González-Fuente et al., 2015; Nota et al., 2021; Tomasello et al., 2019). Thus, research in this domain likewise requires (and sometimes already includes; e.g., González-Fuente et al. 2015, Nota et al. 2021) fine-grained annotation of facial expressions and other visual cues in video data.

Annotation of NMMs is highly time-consuming and also poses challenges for data analysis, given the considerable number of possible NMMs and the fact that temporal information is ideally also taken into account. Even so, as we have discussed, such work is vital both for empirical assessment of theoretical claims as well as to gain more insight into the factors that lead to variation in NMMs use in sign language and multimodal communication.

Currently, the field lacks standard guidelines for annotating NMMs. That is to say, guidelines for annotating NMMs do exist, but none have been thoroughly validated and have become a community-wide standard. Researchers studying NMMs often end up devising new annotation protocols tailored to their specific research objectives.<sup>3</sup> Furthermore, we also lack a standard method to quantify inter-annotator agreement. In fact, publications in sign language linguistics rarely report inter-rater agreement scores. For instance, ten out of the seventeen research articles published in *Sign Language & Linguistics* in 2021-2023 investigate properties of sign languages based on annotated video data, but just one of them reports inter-annotator agreement scores. Adopting a standard method for this purpose would benefit the field by increasing data transparency, and would enable us to iteratively evaluate and improve our annotation guidelines.

The general project that the present paper is part of therefore pursues (i) the development of a reliable protocol for the annotation of NMMs, and (ii) a procedure for evaluating inter-annotator agreement. This paper focuses on the second project pillar. Indeed, it does not really matter for the purpose of this paper which annotation protocol we evaluate.

---

<sup>3</sup>A reviewer made us aware of an extensive annotation protocol for both manual and non-manual markers that was developed in the context of the SignStream project (Neidle, 2002). While this annotation scheme has to our knowledge not been evaluated for inter-annotator agreement, some of the general and specific insights and recommendations discussed in these guidelines overlap with those discussed in the present paper. We thank the reviewer for pointing us to this work, and we will briefly return to it in our discussion on the distinction between *poses* and *movements* in Section 5.1.

### 3. Summary of Oomen et al. (2023)

In Oomen et al. (2023) we presented a first version of the annotation guidelines, according to which two coders annotated a test set of 60 interrogative sentences in Sign Language of the Netherlands (NGT), which came from a larger dataset created in the context of another study. The annotations were produced in ELAN (2023). Coder 1 (C1) annotated 585 events over the 12 tiers specified in the guidelines, and Coder 2 (C2) annotated 564 events. The tiers concerned the eyebrows, eye shape, eye gaze direction, shoulder position, body position, head position and movement, mouth configuration, lip corner configuration, and nose wrinkle.

In Oomen et al. (2023), we already briefly evaluated the reliability of the resulting annotations and included a few recommendations for the improvement of the annotation guidelines. However, the discussion was limited to one annotation tier (concerning the eyebrows) and one evaluation method. In the present paper, we provide a more in-depth evaluation, and offer more extensive recommendations to improve the guidelines.

### 4. Evaluation Methods and Measures

Video-recorded sign language data represents so-called *timed-event sequential data* (Bakeman et al., 2009; Bakeman and Quera, 2011). In general, such data involve recordings of sequences of events, each with a particular time duration. Besides sign linguists, researchers investigating other phenomena (e.g., speech, multimodal communication, or animal behavior) also work with this kind of data, make similar use of annotations, and have devised several methods to assess inter-annotator agreement for this type of data. Broadly, two approaches can be distinguished: *frame-based* approaches and *event-based* approaches (Bakeman et al., 2009).<sup>4</sup> In both these approaches, inter-annotator agreement is quantified using *confusion matrices* and *agreement indices*. We briefly explain each of these methods in this section.

#### 4.1. The Event-Based Approach

In the event-based approach, we treat all annotations as ‘events’, and first determine the temporal overlap between annotations of the two coders, who we refer to as C1 and C2. This is done separately for each tier. For this approach, the annotation label ‘neutral’ (used when a particular facial feature or body part is in a neutral position) is not classified as an event, so these labels are disregarded. Two annotations are taken to ‘match’ if their

---

<sup>4</sup>Frame-based approaches are also referred to as *time-based* approaches (Bakeman et al., 2009).



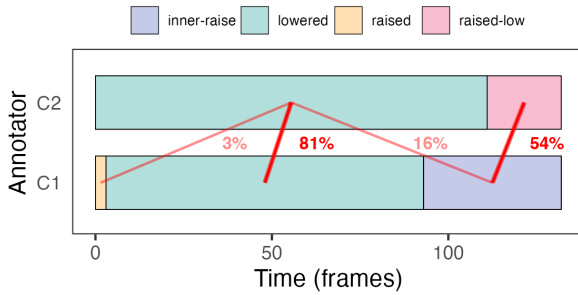


Figure 1: Annotations for the eyebrow tier of a sentence. Red lines show the percentage overlap between all annotations; the thick lines show the percentage overlap between ‘matching’ annotations.

overlap exceeds a pre-defined overlap threshold. At this stage, the label *values* are not considered: matches are established purely based on temporal overlap. We use an overlap threshold of 51%. Overlap between two annotations  $i$  and  $j$  is calculated according to the following formula (Holle and Rein, 2015):

$$O_{ij} := \frac{\min(\text{offset}_i, \text{offset}_j) - \max(\text{onset}_i, \text{onset}_j)}{\max((\text{offset}_i - \text{onset}_i), (\text{offset}_j - \text{onset}_j))}$$

In words,  $O_{ij}$  is the length of the overlap between  $i$  and  $j$  divided by the length of the longest of the two annotations. If  $O_{ij}$  does not exceed the threshold,  $i$  and  $j$  are not regarded as a match. If an annotation by C1 does not have any matching annotations by C2, that annotation is regarded as ‘unmatched’.

Figure 1 shows the annotations by C1 and C2 for the eyebrow tier of an example sentence in our test dataset. The red lines show the percentage overlap between all annotations of C1 and C2, respectively. The thin transparent lines show the percentage overlap between ‘unmatched’ annotations, while the ‘matching’ annotations are illustrated by the thick opaque lines. Again, note that ‘matching’ annotations do not necessarily involve the same label, the only criterion is that they have sufficient temporal overlap. We turn to quantifying the extent to which matching annotations agree in terms of their labels in Section 4.3.

## 4.2. The Frame-Based Approach

On the frame-based approach, we simply consider each individual frame in all videos annotated by C1 and C2, and then determine whether the labels applied by C1 and C2 to each of these frames correspond. We do this separately for each tier. On this approach we do take ‘neutral’ labels into account, so that for each frame we can compare the labels that the two coders assigned.

## 4.3. Confusion Matrices

Both on the event-based approach and on the frame-based approach, the first step in quantifying inter-annotator agreement is to compile a so-called *confusion matrix*. For examples of confusion matrices for two of the tiers we evaluated, see Section 5.2 and 5.3. Cell  $ij$  in a confusion matrix displays the number or the percentage of events/frames which C1 labeled as  $i$  and C2 labeled as  $j$ . When displaying percentages, a confusion matrix is either constructed from the perspective of C1 (which means that all rows add up to 100%) or from the perspective of C2 (all columns add up to 100%).

## 4.4. Agreement Indices

Besides confusion matrices, another way to quantify inter-annotator agreement is to compute *agreement indices* for each label. Here it is important to note that so-called *raw* agreement indices are insufficient. To illustrate this, suppose that two annotators  $x$  and  $y$  label 100 items. To 50 items they both apply label A, to 20 items only  $x$  applies label A, to 20 items only  $y$  applies label A, and to the final 10 items they both apply another label. Then,  $x$  and  $y$  agree in  $50 + 10 = 60$  of the cases as to whether label A applies or not. The raw agreement index for label A, then, is 0.6. However, this does not take into account the possibility that, at least in some cases,  $x$  and  $y$  may have agreed on the application of label A *by mere chance*. Both  $x$  and  $y$  applied label A to 70% of the items, and other labels to 30% of the items. If they would randomly assign label A to 70% and other labels to 30% of the items, they would agree 58% of the time as to whether A applies or not (because  $(0.7 * 0.7) + (0.3 * 0.3) = 0.58$ ). So the raw agreement index,  $i_{raw} = 0.6$ , is just slightly higher in this case than the chance agreement index,  $i_{chance} = 0.58$ . Chance-corrected agreement indices take this factor into account.

One widely used chance-corrected index is Cohen’s  $\kappa$  (Cohen, 1960). It is computed by dividing the difference between  $i_{raw}$  and  $i_{chance}$  by the difference between  $i_{chance}$  and the index for perfect agreement, which is 1.

$$\kappa := (i_{raw} - i_{chance}) / (1 - i_{chance})$$

In the example above,  $\kappa$  would amount to  $0.02 / 0.42 = 0.05$ . To give some other examples, if  $i_{raw} = 0.7$  and  $i_{chance} = 0.5$  then  $\kappa = 0.4$ , and if  $i_{raw} = 0.9$  and  $i_{chance} = 0.6$  then  $\kappa = 0.75$ .

It is important to note that it is not straightforward to interpret agreement indices such as Cohen’s  $\kappa$ . Some researchers have proposed specific interpretations. For instance, a frequently cited interpretation is that of Landis and Koch (1977, 165), who posit that a  $\kappa$  score of 0.21–0.40 amounts to ‘fair’ agreement, 0.41–0.60 to ‘moderate’ agreement, 0.61–0.80 to ‘substantial’ agreement, and

0.81–1 to ‘almost perfect’ agreement. However, it has been noted in the literature that such absolute interpretations are arbitrary and problematic, because  $\kappa$  scores can be affected by *label prevalence* (whether the labels are equiprobable or not), *coder bias* (whether the marginal probabilities for the two coders are similar or different), and the *number of possible labels* for a given annotation tier (Bakeman et al., 1997; Sim and Wright, 2005).

Thus, not too much should be read into any single  $\kappa$  score on its own. Rather, a  $\kappa$  score should always be considered *relative to other  $\kappa$  scores*. For instance, if there are three roughly equiprobable labels for a given annotation tier (A, B, C), and the  $\kappa$  score for A is much lower than that for B and C, then we can conclude that the instructions for label A in the annotation guidelines were less reliable than those for B and C. Another possibility is to compare  $\kappa$  scores across iterations of the annotation guidelines. With every new iteration, we hope to obtain higher  $\kappa$  scores. If we do, this confirms that the adjustments we made indeed succeeded in making the protocol more reliable. The latter type of comparison is our main intended use of  $\kappa$  scores. That is, we mainly report  $\kappa$  scores here for comparison with future iterations of the guidelines.<sup>5</sup>

## 5. Results and Recommendations

We have compiled confusion matrices and  $\kappa$  scores for all twelve tiers in the annotation guidelines, based on the test dataset from Oomen et al. (2023) described above, both under the event-based approach and under the frame-based approach. Based on our analysis and comparison of these twelve tiers, we formulate a number of general recommendations for improvement of the annotation guidelines in Section 5.1. For reasons of space, we cannot discuss the results for all tiers individually; they are presented in a technical report which is available in the supplementary materials. Here, we only discuss two specific tiers, *head y* (with labels ‘up’, ‘down’ and ‘neutral’; Section 5.2) and *head move* (with labels ‘nod’, ‘nodding’, ‘shake’, ‘shaking’, ‘sideways’, and ‘neutral’; Section 5.3), as they

---

<sup>5</sup>The event-based method of Holle and Rein (2015) that we have described in this section is implemented in ELAN and can be performed straightforwardly by selecting File → Multiple File Processing → Calculate Inter-Annotator Reliability. The output is a .txt file with agreement matrices and Cohen’s  $\kappa$ . We have re-implemented the method in R with additional visualisation functionalities (see Section 9 for a link to the documented R script). Advantages of the R script over the ELAN functionality are (i) that it is fully transparent and (ii) that it can easily be modified and extended (see Section 6 for some suggestions in this direction), and (iii) that the results can be visualised in various ways.

relate to many issues that we target with our general recommendations.

### 5.1. General Recommendations

The most important general insight we obtained is that a methodical distinction should be made between two types of NMM, which we refer to as *poses* and *movements*. As a reviewer pointed out, a similar distinction is made in the SignStream annotation protocol (Neidle, 2002), namely a distinction between ‘positions’ and ‘movements’. The former involve some part of the face or body ‘first moving to a target position and then maintaining that position’ for some time, while the latter involve ‘continuous (potentially repeated) movements’ (Neidle, 2002, p.24).

Very much in line with this, we define a *pose* as a non-manual feature which can be characterized in terms of a *single configuration* of part of the face or body, which is *held* for a certain amount of time. Disregarding transitional movements in and out of a pose (see below for discussion on how to treat such transitions), a pose itself does not involve inherent movement. Clear examples of poses are the features ‘head up’ and ‘head down’ on the *head y* tier (see Section 5.2). Poses can in principle be labeled on a frame-by-frame basis.

On the other hand, we define *movements* as non-manual features for which a *temporal progression* from a certain starting configuration, possibly through certain intermediate configurations, to a certain target configuration is characteristic. Movements typically happen within a relatively short amount of time. Many movements are oscillatory; in this case the target configuration is the same as the starting configuration. Clear examples of movement NMMs are head nods and headshakes on the *head move* tier (see Section 5.3), and eye blinks. Since movements cannot be characterized in terms of a single configuration but involve a temporal progression through multiple configurations, they can never be identified based on a single video frame only. Labeling a video segment as involving a certain movement is thus qualitatively different from labeling it as involving a certain pose, as the entire sequence of frames within the given segment—and not each frame individually—determines the annotation value.<sup>6</sup>

This discussion yields three concrete recommendations that should be integrated in future versions of the annotation guidelines.

Firstly, in the current version of the annotation guidelines, certain tiers contain labels for both *poses* and *movements*, as exemplified by the *head move* tier discussed in Section 5.3. Given the

---

<sup>6</sup>An analogy: movement labels are like *collective predicates*, while pose labels are like *distributive predicates*.

qualitative differences between poses and movements that we just identified, annotation tiers should comprise either poses or movements, not both. It should also be made explicit for each annotation label whether it describes a pose or a movement. This is lacking in the current guidelines, and it is evident that this sometimes led to confusion among coders. For instance, the label ‘closed’ on the *eye gaze* tier was applied differently by our coders. One coder used it only to label longer segments where the signer kept their eyes closed (an *eye pose*). The other coder used the label in such cases too, but also applied it to short eye blinks (an *eye movement*). Section 5.2 discusses another example.

Secondly, on *pose* tiers, both neutral and non-neutral configurations (e.g. ‘head neutral’ vs. ‘head up’ or ‘head down’) should be annotated, because neutral configurations are poses as well. As a consequence, *pose* tiers are typically *continuous*, in the sense that every video segment is given some label.<sup>7,8</sup> In contrast, on *movement* tiers, only movement events should be annotated; if there is no movement that corresponds to one of the labels on the tier, nothing should be annotated. For example, on a tier for eye blinks, each blink should be labeled, but no further annotations should be added; ‘neutral’ is not a useful label in this case since it does not describe a movement.

Finally, the guidelines should specify what it means for a *pose* to be held “for a certain amount of time”, and for a *movement* to occur “within a relatively short amount of time”. For instance, if we specify within which time frame a signer’s eyes should close and re-open for it to be considered an *movement*, i.e. a blink, instead of a *pose*, then coders can make a principled distinction between these two labels in situations where there may otherwise be confusion. We plan to undertake empirical work to determine suitable thresholds.

Relatedly, there is the issue of when a *pose* or *movement* should start and end. This issue is particularly tricky when it comes to *poses*: at what point should a coder decide that a signer’s eyebrows are no longer in, say, a ‘neutral’ position, but have rather become ‘raised’? As a basic principle, we propose that pose annotations should include the transition movement *into* the pose but not the one *out of* that pose (and into the next one).<sup>9</sup>

---

<sup>7</sup>There are exceptions to this. For instance, on the *pose* tier for eye gaze direction, segments in which the eyes are closed need not be given a label.

<sup>8</sup>What we suggest here for poses differs from the treatment of ‘positions’ in the SignStream protocol; neutral positions are not regarded there as true positions and as such are not annotated.

<sup>9</sup>This differs, again, from the SignStream protocol, where transition movements in and out of positions are coded separately, as ‘s(tart)’ and ‘e(nd)’, respectively.

Another important insight we obtained concerns tier structure. With twelve tiers, the current guidelines already contain a fairly elaborate tier structure, yet we found that further distinctions between tiers and/or annotation labels are desired for reasons of clarity, exhaustiveness, and systematicity. Moreover, an extensive tier structure makes it easier for researchers to focus on only specific NMMs. We therefore propose the following principles for systematic expansion of the tier structure: (1) Every tier should concern a **UNIQUE BODY PART** (e.g. head, eyelids, nose, eyebrows); (2) Every tier should only include labels for *poses*, or only for *movements* (e.g. the eyelid *movement* ‘blink’ should be annotated on a different tier than the eyelid *pose* ‘closed’); (3) Every tier should contain labels that are **MUTUALLY EXCLUSIVE** (i.e., any two NMMs that can co-occur should be annotated on separate tiers); (4) The set of labels for *pose* tiers should be **JOINTLY EXHAUSTIVE** – i.e., each *pose* tier should have a set of labels that cover the full range of possible *poses* for the relevant body part (as discussed above, this does not apply to *movement* tiers); (5) The set of labels on a given tier should be sufficiently **CONTRASTIVE**.

Regarding criterion (5), some tiers in the current guidelines include pairs of labels that describe the same NMM but to different degrees of engagement (e.g., ‘squint-full’ and ‘squint-half’ on the ‘eye shape’ tier). Our analyses show that the inclusion of such labels generally lead to poor inter-coder agreement. We suggest to only include the label ‘squint’ in future versions of the guidelines.

While it seems impossible to reliably annotate the degree of engagement of non-manual features, we do believe it is useful to obtain a measure of the *confidence level* of the coders (previously explored, for instance, for annotation of emotions in text by Troiano et al. 2021). Coders may record, for every annotation event, their level of confidence in the label they applied, on a three-point scale from low to high. Researchers then have the option to only analyze a subset of the data with high confidence scores, and to compare this analysis to one taking the entire dataset into account. Moreover, confidence ratings would be useful as training data for machine learning in the future.

In such a system, including ‘neutral’ poses in the repertoire of possible poses is important. Say a study only wishes to include annotations with high confidence ratings, but ‘neutral’ poses are not labeled to begin with. Then for all events that are not considered, it is unknown whether they are not included because they received a low confidence rating or because they involve a neutral state.

Besides a sub-tier for confidence ratings, another sub-tier we propose to add is one on which annotators can indicate when a particular non-manual feature clearly does not have a communicative func-



tion, e.g. when a signer wrinkles their nose because it's itching, or turns their head because of an unexpected movement next to them. In such cases, coders can make a note on this tier, allowing for irrelevant events to be excluded from the analysis.

Furthermore, poses and especially movements should be illustrated in the guidelines not just with static video stills but also with video clips or GIFs. As such, the next version of the guidelines should be constructed in digital format such as in the form of a website or a slide deck.

A final recommendation does not concern the guidelines, but rather the data collection method. A major challenge that arises when manually annotating video data is that it involves analyzing 2D data that represents a 3D reality. Specifically, we found that a single (near-)frontal camera view makes the work for manual coders particularly challenging. We therefore advise researchers collecting data to always use multiple cameras, including a side-view camera. In addition, 3D capturing techniques may be considered as well (see [Esselink et al., 2023](#)).

## 5.2. *Head y*

On the *head y* tier (a *pose* tier) there were three possible labels: 'up', 'down', and 'neutral'.

**Frame-Based Approach** The confusion matrices in Table 1 show that the coders generally agreed on the 'neutral' label, but not on 'down' and 'up'.

**Event-Based Approach** The event-based confusion matrices in Table 2 show that the two coders identified a similar number of events as 'down' or 'up' events. However, the agreement rates concerning these events are extremely low. In total, only 15% of the 68 events annotated on this tier matched another event with the same label.

**Error Analysis** To better understand the low agreement scores for this tier, we carried out an error analysis of the mismatched events. We found that 3/19 [3/23] unmatched events labeled as 'down' by C1 [C2] were unmatched due to the coders not agreeing on onset and/or offset, resulting in insufficient overlap between the events to establish a match. For 2/19 [4/23] events, C1 [C2] had labeled (almost) the entire sentence as 'down', but C2 [C1] labeled two short events as 'down', which were preceded and followed by 'neutral' interludes. For the remaining 14/19 [16/23] unmatched events, C1 [C2] had identified (usually quite short) parts of the sentence as 'down' events, whereas C2 [C1] labeled these segments as 'neutral'.

For all unmatched 'up' events, one of the coders labeled the relevant segment as 'neutral'.

**Cohen's Kappa** On the frame-based approach, the  $\kappa$  scores are very low: 0.27 ('down'), 0.27 ('up'), and 0.21 ('neutral'). On the event-based approach, they are even worse: -0.27 ('down') and 0.14 ('up').

**Tier-specific Recommendations** The results for the *head y* tier show that the coders were hardly consistent with each other in identifying 'up' and 'down' events. In most cases, the disagreements were *categorical*, i.e., one coder identified an 'up' or 'down' event while the other coder labeled the same segment as 'neutral'.

Based on these results, we have three specific recommendations for this tier. First, we expect that use of a second camera offering a side view would facilitate more accurate and consistent coding of head position. Second, the annotation guidelines need to be more explicit on how much the head should diverge from a neutral position in order for it to count as a head 'up' or 'down' event. And third, the guidelines should specify a minimum duration of 'up' and 'down' events, in particular so as to distinguish 'down' events from *head nods* (see Section 5.3 below). In future work, we aim to establish concrete minimum duration values to be included in the guidelines.

## 5.3. *Head move*

The *head move* tier is intended for annotating head *movements*, and includes the labels 'nod' (single nod), 'nodding' (multiple nods), 'shake' (single shake), 'shaking' (multiple shakes), 'sideways' (single sideways movement of the head), and 'neutral'.

**Frame-Based Approach** For this tier, there is generally not much confusion between the coders. One might have expected low agreement on the labels 'nod' vs 'nodding', and 'shake' vs 'shaking', but Table 3 shows that this is not necessarily the case. However, we can make some other interesting observations pertaining to these labels.

Overall, C2 applied the various labels (other than 'neutral') to more frames than C1, who used 'neutral' more often. An especially interesting pattern can be observed for the label 'nod': when C1 applied this label, C2 agreed 52% of the time, labeling the remaining frames as 'nodding' (23%) or 'neutral' (25%). When C2 used 'nod', C1 only agreed 26% of the time. The remaining 74% of frames were labeled overwhelmingly as 'neutral' (69%). Both coders applied 'nodding' quite similarly, although C2 again labeled more frames as such than C1.

For 'shake' and 'shaking', we see a large disparity in application for both coders. The label 'shake' is barely assigned to any frames, totalling only 87 frames for C1, and 57 frames for C2. In contrast, the label 'shaking' is applied to a large number



Table 1: Confusion matrix for the *head y* tier showing the total number of frames (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of frames					(b) C1				(c) C2				
C1/C2	down	up	neutral	Total	C1/C2	do	up	ne	Total	C1/C2	do	up	ne
down	597	24	1086	1707	do	<b>35</b>	1	64	100	do	<b>45</b>	6	17
up	6	102	165	273	up	2	<b>37</b>	61	100	up	0	<b>26</b>	3
neutral	720	273	4920	5913	ne	12	5	<b>83</b>	100	ne	55	68	<b>80</b>
Total	1323	399	6171	7893	Total	100	100	100		Total	100	100	100

Table 2: Confusion matrix for the *head y* tier showing the total number of events (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of events					(b) C1				(c) C2				
C1/C2	down	up	unmatched	Total	C1/C2	do	up	un	Total	C1/C2	do	up	un
down	7	0	19	26	do	<b>27</b>	0	73	100	do	<b>23</b>	0	76
up	0	3	6	9	up	0	<b>33</b>	67	100	up	0	<b>23</b>	24
unmatched	23	10	0	33	un	70	30	0	100	un	77	77	0
Total	30	13	25	68	Total	100	100	100		Total	100	100	100

of frames, totalling 1704 frames for C1, and 1926 for C2. Again, we see a similar pattern as above, where C2 assigned this label to more frames than C1, who mostly labeled these remaining frames as ‘neutral’. However, in this case there is a higher level of agreement: C2 agreed with the ‘shaking’ labels applied by C1 99% of the time, and C1 agreed with C2 88% of the time.

Finally, C2 applied the label ‘sideways’ to 144 frames, which were all labeled as ‘neutral’ by C1. C1 never applied the label ‘sideways’.

**Event-Based Approach** The confusion matrices in Table 4 for the event-based approach show the same general patterns as the confusion matrices of the frame-based approach in Table 3. There is barely any confusion between the labels ‘nod’/‘nodding’ and no confusion between the labels ‘shake’/‘shaking’. Looking closer at the data, we see that the confusion between these labels for the frame-based approach can be mostly attributed to disagreement on the onsets and offsets of events.

The labels ‘nodding’, ‘shake’, and ‘shaking’ were applied to a similar number of events by both coders, with the total number of events assigned one of these labels differing by only 1. This shows that, as C2 generally applied these label to more frames than C1, the annotation events by C2 were likely longer in duration than those of C1. For the label ‘nod’, we see a big disparity in the number of annotation events: C1 labeled 15 events as such, while C2 assigned this label to 26 events. The majority of these events were unmatched for both C1 and C2. The labels ‘shake’ and ‘sideways’ were barely assigned to any events by the coders.

**Error Analysis** A possible explanation for the disparity between the frames and events labeled as ‘nod’ by C2 and as ‘neutral’ by C1 is that C1 labeled these instances as ‘down’ (in the *head y* tier) instead. We briefly examine this possibility here; Table 5 shows the events of interest. In 19 cases, C2 labeled an event on the *head move* tier as ‘nod’ while C1 labels it as ‘neutral’. We examine labels given to corresponding events in the *head y* tier. The rows display the labels given to these events by C1; the columns display the labels given to the matching event by C2.

In 9 cases, C1 labels a corresponding event on the *head y* tier as ‘down’; of these 9 cases, C2 labels the corresponding event as ‘down’ twice, and as ‘neutral’ 7 times. However, also in 9 cases, C1 labels a corresponding event on the *head y* tier as ‘neutral’; of these, C2 labels the corresponding event as ‘down’ 3 times, and as ‘neutral’ 6 times. In one case, C1 labels the corresponding event as ‘up’, while C2 labels this event as ‘neutral’.

Therefore, we see that in about 50% of the cases examined here, C1 labeled the events as ‘down’ on the *head y* tier instead of ‘nod’ on the *head move* tier. We cannot definitively conclude that in these cases, C1 labeled events as ‘down’ in the *head y* tier in lieu of labeling the corresponding events as ‘nod’ in the *head move* tier. However, this does explain some of the discrepancy.

This leads us to another interesting observation. C2 labeled 5 events as ‘nod’ in the *head move* tier, as well as labeling a simultaneous event as ‘down’ in the *head y* tier. We find that for 4 of these occurrences, the events on both tiers have roughly the same onsets and offsets. A quick check of the annotations provided by C1 reveals 4 ‘nod’ events

Table 3: Confusion matrix for the *head move* tier showing the total number of frames (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of frames							
C1/C2	nod	nodding	shake	shaking	sideways	neutral	Total
nod	183	81	0	0	0	90	354
nodding	27	567	0	3	0	60	657
shake	0	0	51	21	0	15	87
shaking	6	0	0	1686	0	12	1704
sideways	0	0	0	0	0	0	0
neutral	489	240	6	216	144	3996	5091
Total	705	888	57	1926	144	4173	7893

(b) C1								(c) C2						
C1/C2	nd	ng	se	sg	si	ne	Total	C1/C2	nd	ng	se	sg	si	ne
nd	<b>52</b>	23	0	0	0	25	100	nd	<b>26</b>	9	0	0	0	2
ng	4	<b>86</b>	0	0	0	9	100	ng	4	<b>64</b>	0	0	0	1
se	0	0	<b>59</b>	24	0	17	100	se	0	0	<b>89</b>	1	0	0
sg	0	0	0	<b>99</b>	0	1	100	sg	1	0	0	<b>88</b>	0	0
si	0	0	0	0	<b>0</b>	0	0	si	0	0	0	0	<b>0</b>	0
ne	10	5	0	4	3	<b>78</b>	100	ne	69	27	11	11	100	<b>96</b>
								Total	100	100	100	100	100	100

Table 4: Confusion matrix for the *head move* tier showing the total number of events (a) and the percentage-wise confusion matrices from the perspective of C1 (b) and C2 (c)

(a) Total number of events							
C1/C2	nod	nodding	shake	shaking	sideways	unmatched	Total
nod	6	0	0	0	0	9	15
nodding	1	9	0	0	0	4	14
shake	0	0	3	0	0	1	4
shaking	0	0	0	19	0	3	22
sideways	0	0	0	0	0	0	0
unmatched	19	4	0	4	5	0	32
Total	26	13	3	23	5	17	87

(b) C1								(c) C2						
C1/C2	nd	ng	se	sg	si	un	Total	C1/C2	nd	ng	se	sg	si	un
nd	<b>40</b>	0	0	0	0	60	100	nd	<b>23</b>	0	0	0	0	53
ng	7	<b>64</b>	0	0	0	29	100	ng	4	<b>69</b>	0	0	0	24
se	0	0	<b>75</b>	0	0	25	100	se	0	0	<b>100</b>	0	0	6
sg	0	0	0	<b>86</b>	0	14	100	sg	0	0	0	<b>83</b>	0	18
si	0	0	0	0	<b>0</b>	0	0	si	0	0	0	0	<b>0</b>	0
un	60	12	0	12	16	<b>0</b>	100	un	73	31	0	17	100	<b>0</b>
								Total	100	100	100	100	100	100

in the *head move* tier with simultaneous events in the *head y* tier labeled as ‘down’ or ‘up’. However, the onset and offset of events in these tiers do not match up, meaning that the head was angled as either ‘down’ or ‘up’ for a longer period of time, within which a ‘nod’ took place. We can conclude that C1 did not confuse the meaning of ‘nod’ on

the *head move* tier and ‘down’ on the *head y* tier, whereas the difference between these labels was not always clear for C2.

**Cohen’s Kappa** For the frame-based approach, the  $\kappa$  indices for ‘nodding’ (0.71), ‘shake’ (0.71), and ‘shaking’ (0.91) are reasonably high, as expected.

C1/C2	down	neutral	up	Total
down	2	7	0	9
neutral	3	6	0	9
up	0	1	0	1
Total	5	14	0	19

Table 5: Labels given to events in the *head y* tier, occurring simultaneously with events in the *head move* tier, which have been labeled as ‘neutral’ by C1, and ‘nod’ by C2

The  $\kappa$  index for ‘nod’ (0.30) is much lower, as there was a lot of disagreement about this label between the coders. The  $\kappa$  index for sideways is 0.00, as the coders never agreed on this label.

The  $\kappa$  indices for labels in the event-based approach are generally lower than those of the frame-based approach. However, the indices are still relatively high for ‘nodding’ (0.61), ‘shake’ (0.85), and ‘shaking’ (0.79). The index for ‘nod’ is lowered to 0.09, while the index for ‘sideways’ remains 0.00.

**Tier-specific Recommendations** Firstly, we note that all labels on the *head move* tier can be categorized as (oscillating) *movements*, with the exception of ‘sideways’, which is a *pose*. The latter should therefore be moved to a separate *pose* tier.

Secondly, although head nods and headshakes involve the same body part, are mutually exclusive, and contrastive (see Section 5.1), we recommend that head nods and headshakes are annotated on separate tiers because they serve very different functions in sign languages. This way, researchers interested only in headshakes need not annotate head nods and vice versa.

Finally, the annotation guidelines should include clear descriptions of what constitutes a ‘nod’ (*movement*) and a head ‘down’ (*pose*), with concrete temporal indications for the required length of *movements* vs. *poses* (in terms of time rather than frames, as users may use different frame-rates). The guidelines should warn that these features can look similar, and show examples of the differences between them.

## 6. Discussion of Evaluation Methods

Considering the assessment of inter-annotator agreement for timed-event sequential data in general, Bakeman et al. (2009, 146) advise the use of both event-based and frame-based methods, as “each provides somewhat different . . . but valuable information as to how observers are disagreeing, and are thus useful in different ways as observers strive to improve their agreement”. We will now briefly discuss some concrete benefits of these methods we identified for NMM data.

An advantage of the event-based approach is that it allows for an error analysis, as illustrated in Section 5.2. This error analysis goes beyond confusion matrices and  $\kappa$  scores: each unmatched event can be examined to determine the *types* of errors that caused the mismatches. This information helps determine which concrete changes to the annotation guidelines would be most effective.

Turning to the frame-based approach, the main purpose for our use-case is that—in combination with the event-based approach—it provides an indication of the nature of the disagreements between coders. In particular, if the frame-based approach yields higher agreement scores than the event-based approach, this suggests that the low agreement scores on the event-based approach are partly due to the following type of mismatches. Say C1 coded an entire sentence as ‘down’ on the *head y* tier, while C2 coded three separate long segments within that sentence as ‘down’, interspersed with two short ‘neutral’ segments. With the event-based method, all events coded on this tier would be regarded as unmatched. The frame-based method, on the other hand, would only count the disagreement of the ‘neutral’ segments; the rest would count as agreement.

The combination of the two approaches thus gives a more well-rounded overview of how the coders disagree. The event-based approach serves as a basis, supplemented by the frame-based approach. However, we should note that the frame-based approach, while in some cases providing an indirect indication of how coders disagreed, never provides a definitive insight into this important question.

Therefore, we propose to develop, in future work, an enriched version of the event-based method, which automatically categorizes the error-types of unmatched events (such as in the error analysis in Section 5.2 for the *head y* tier). This method would keep track of additional information such as the duration of the events that the coders agreed and disagreed on, and for each unmatched event, what type of error caused the mismatch. With this enriched event-based approach, the frame-based approach would become superfluous for our use-case, as the enriched event-based approach would provide all the necessary information to further improve the annotation procedure.

## 7. Conclusion

We evaluated guidelines for annotating NMMs by examining a test dataset involving two coders. We used a frame-based and an event-based approach to calculate inter-annotator agreement. Based on the results, we formulated concrete recommendations to further refine the annotation guidelines.

## 8. Acknowledgements

We thank Marc Schulder and James Trujillo for helpful discussion. We also gratefully acknowledge the three reviewers for their valuable feedback. This work is part of the project *Questions in Sign Language* (grant number VI.C.201.014, PI Roelofsen) financed by the Dutch Science Foundation (NWO).

## 9. Supplementary Materials

1. Annotation manual: <https://doi.org/10.21942/uva.24080868>
2. ELAN annotation template: <https://doi.org/10.21942/uva.22732616>
3. Inter-annotator agreement scripts: <https://doi.org/10.21942/uva.24080724>
4. Testset videos: <https://doi.org/10.21942/uva.21666203>
5. Testset annotation files: <https://doi.org/10.21942/uva.22737074>
6. Technical report: <https://doi.org/10.21942/uva.25563540>

## 10. Bibliographical References

- Benjamin Bahan. 1996. *Non-manual realization of agreement in American Sign Language*. Ph.D. thesis, Boston University.
- Roger Bakeman, Duncan McArthur, Vicenç Quera, and Byron F Robinson. 1997. [Detecting sequential patterns and determining their reliability with fallible observers](#). *Psychological Methods*, 2(4):357.
- Roger Bakeman and Vicenç Quera. 2011. Sequential analysis and observational methods for the behavioral sciences.
- Roger Bakeman, Vicenç Quera, and Augusto Gnisci. 2009. [Observer agreement for timed-event sequential data: A comparison of time-based and event-based algorithms](#). *Behavior Research Methods*, 41(1):137–147.
- Janet Bavelas and Nicole Chovil. 2018. [Some pragmatic functions of conversational facial gestures](#). *Gesture*, 17:98–127.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and psychological measurement*, 20(1):37–46.
- ELAN. 2023. Version 6.5. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [\[link\]](#).
- Lyke Esselink, Oomen Marloes, and Roelofsen Floris. 2023. [Exploring new methods for measuring, analyzing, and visualizing facial expressions](#). *FEAST. Formal and Experimental Advances in Sign language Theory*, 5:35–48.
- Aslı Göksel and Meltem Kelepir. 2013. [The phonological and semantic bifurcation of the functions of an articulator: HEAD in questions in Turkish Sign Language](#). *Sign Language & Linguistics*, 16:1–30.
- Santiago González-Fuente, Victoria Escandell-Vidal, and Pilar Prieto. 2015. [Gestural codas pave the way to the understanding of verbal irony](#). *Journal of Pragmatics*, 90:26–47.
- Henning Holle and Robert Rein. 2015. [Easydiag: A tool for easy determination of interrater agreement](#). *Behavior Research Methods*, 47(3):837–847.
- Andrea Lackner. 2017. *Functions of head and body movements in Austrian Sign Language*. De Gruyter Mouton, Berlin.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Carol Neidle. 2002. [Signstream annotation: Conventions used for the American Sign Language linguistic research project](#). Technical report no. 11.
- Naomi Nota, James P. Trujillo, and Judith Holler. 2021. [Facial signals and social actions in multimodal face-to-face interaction](#). *Brain Sciences*, 11:1017.
- Marloes Oomen, Lyke D. Esselink, Tobias de Ronde, and Floris Roelofsen. 2023. [First steps towards a procedure for annotating non-manual markers in sign languages](#). In *NELS 53: Proceedings of the Fifty-Third Annual Meeting of the North East Linguistic Society*, volume 2, pages 257–266. GLSA.
- Nina-Kristin Pendzich. 2020. *Lexical nonmanuals in German Sign Language. Empirical studies and theoretical implications*. De Gruyter Mouton.
- Julius Sim and Chris C Wright. 2005. [The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements](#). *Physical Therapy*, 85(3):257–268.
- Rosario Tomasello, Cora Kim, Felix R. Dreyer, Luigi Grisoni, and Friedemann Pulvermüller. 2019. [Neurophysiological evidence for rapid processing of verbal and gestural information in understanding communicative actions](#). *Scientific Reports*, 9:16285.



Enrica Troiano, Sebastian Padó, and Roman Klinger. 2021. [Emotion ratings: How intensity, annotation confidence and agreements are entangled](#). In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 40–49, Online. Association for Computational Linguistics.

Ronnie B. Wilbur. 2021. Non-manual markers – theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 530–565.

Ronnie B. Wilbur and Cynthia G. Patschke. 1998. [Body leans and the marking of contrast in American Sign Language](#). *Journal of Pragmatics*, 30:275–303.

# A software editor for the AZVD graphical Sign Language representation system

Michael Filhol , Thomas von Ascheberg

CNRS, LISN, Université Paris–Saclay

Orsay, France

michael.filhol@cnrs.fr, ascheberg@lisn.fr

## Abstract

Based on real spontaneous productions by signers, AZVD is a graphical Sign Language representation system designed to maximise its potential for adoption by the signing community. Additionally, it is kept entirely synthesisable by construction, i.e. any AZVD content determines a signed output, which can be rendered through an avatar for example. This paper reports on the implementation of a software prototype developed to support AZVD editing, and the current extent of AZVD graphics integration. The point is to allow users to experience and discuss the AZVD approach, and ultimately assess it as a standardised graphical form for Sign Language representation.

**Keywords:** Sign Language, graphical form, writing system, AZVD

## 1. Introduction

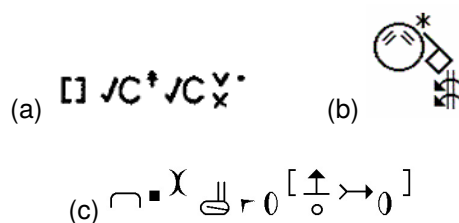
Languages that have a written form are often equipped with software assisting in various types of processing, the first of which being text editing. In contrast, sign languages (SLs) have no written form. While video is often used as a default substitute, it cannot be considered equivalent: its storage is heavy, and it is comparatively laborious to edit, index or query. Moreover, interpreting any of its contents is subject to real time, whereas reading allows to scan and capture multiple parts of the input freely. It also prevents anonymity, which is a significant limitation when considering information and opinion circulation on the internet for example.

First we show a few systems proposed and techniques used by SL users to work around this problem. Then we present the recent “AZee Verbalising Diagram” (AZVD) approach to graphical SL representation, designed to be synthesisable by signing avatars and maximise adoptability by the users. We follow by describing a software editing prototype that we developed to test the system and ultimately evaluate it.

## 2. Verbalising diagrams

To work around or address the lack of adopted SL writing system, some scripts were developed, three of them shown in fig. 1. Some were created for scripting purposes, for linguistic annotation or computer synthesis. Some have claimed a writing system status or potential. But none is adopted by the wide communities of language users (Grushkin, 2017; Kato, 2008).

And yet, there are clues that the need for some form of writing exists. Deaf people and translators



(a) Stokoe's notation (Stokoe et al., 1965)

(b) Sign Writing (Sutton, 2014)

(c) HamNoSys (Prillwitz et al., 1989; Hanke, 2004)

Figure 1: Examples of graphical systems designed for sign languages

also take notes or prepare SL discourses by drawing diagrams that somehow capture their structure, meaning or content in some more or less readable form (Athané, 2015). These diagrams exhibit various arrangements of icons, text, drawings, lines and arrows. An example of such “verbalising diagram” (VD), from the corpus built by Filhol (2020a), is given in fig. 2. It represents an LSF production of 56 s, signed after the diagram was drawn, with the following meaning:

Atoms are very small particles, composed of a nucleus and electrons (elementary negative electric charges). Atoms are electrically neutral because their nucleus holds as many positive charges as electrons do negative charges. Groups of atoms are called molecules. Ions are atoms or molecules with electrons gained or lost from the action of neighbouring atoms. Ions are therefore electrically charged.

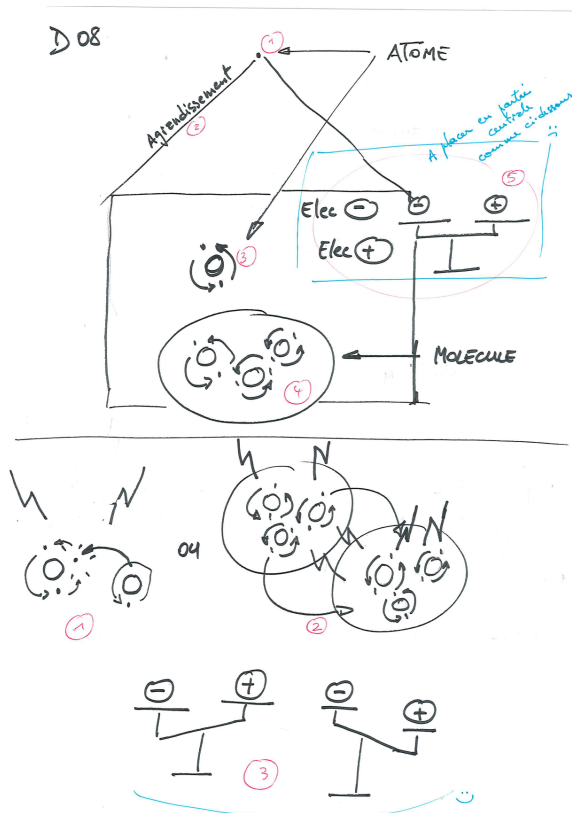
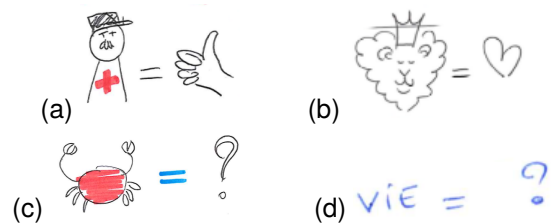


Figure 2: Example of verbalising diagram

VDs are spontaneous productions in the sense that the contained graphics follow no predefined set of rules. This usually makes parts of them readable only by the original author. In other words VDs do not strictly determine the signed form to produce to read them out, so they cannot be viewed as synthesisable input to, say, signing avatars. This is in contrast with a shared property of standardised writing systems, which we consider powerful as content becomes exchangeable in an anonymous, light-weight and editable fashion.

However, after collecting a corpus of VDs from French Sign Language (LSF) users, regularities have been reported both across diagrams and across authors (Filhol, 2020a), to the point where some VD layouts or icons with an identifiable meaning have a systematic signed equivalent when read out by their author. An example is given in fig. 3, where the same '=' symbol is consistently used between a left- and a right-hand side—say  $L$  and  $R$ —to mean that  $R$  is a state or property of  $L$ . This is almost systematically signed with  $L$  and  $R$  in this order with a form of assertion. Another, more trivial example is also visible in the same figure: the '?' symbols here consistently stand for the sign commonly glossed "QUOI" (French for "what").

The spontaneity of the VD representations and the presence of regularities already in the productions led us to propose that a standardised graph-



- (a) Fidel Castro's health = good (F. C. is well)
- (b) lion = nice (the lion is nice)
- (c) cancer = '?' (what is cancer?)
- (d) life = '?' (what is life?)

Figure 3: VD exemplars using the "equal" sign (with meaning in context)

ical script inspired by them could be experienced as a more natural way of representing signed content, hence increase its *adoptability* (Filhol, 2020b). Such a script would include the regular VD layouts when observed, while completing the set for language coverage in a way that it remains *synthesisable*.

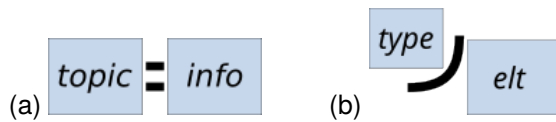
The recent AZVD proposition is a first attempt at satisfying these two features. The next section presents it in further depth.

### 3. AZVD

AZVD is a formal graphical system combining 2D symbols, borrowing from the observed spontaneous ones like that in fig. 3. Similarly to the mathematical script in which atomic tokens (e.g. numbers, variable names) and operators (e.g. unary '!', binary '+', ternary 'Σ') recursively combine to grow formulae of arbitrary size, AZVD allows to build recursive diagrams to represent SL utterances of arbitrary size. To make diagrams synthesisable, every symbol or layout defined in the graphical system is *mapped* to a signed output in a given language, making use of the nested arguments as appropriate. The specification of this output is done with AZee expressions or templates.

AZee is a formal SL representation system used for synthesis with avatars. It defines the notion of *production rule*, i.e. a strong association between a meaning and an articulated form in a SL. The set of production rules for a language is called its *production set*. AZee has already proven efficient in terms of language coverage (Challant and Filhol, 2022) and feasibility and quality of synthesis (McDonald and Filhol, 2021) in LSF.

Some of the regular VD patterns directly correspond to LSF production rules. For example the semantic relationship between elements  $L$  and  $R$  carried by the "equal sign" layout (fig. 3) is exactly the meaning carried by AZee expression  $\text{info-about}(\text{topic}=L, \text{info}=R)$ . An AZVD map-



(a) info-about (*topic*, *info*)  
 (b) instance-of (*type*, *elt*)

Figure 4: AZVD layouts with variable sections, and their AZee expression mappings

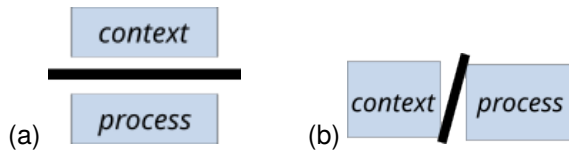


Figure 5: Two AZVD layout variants mapping to in-context (*context*, *process*)

ping is therefore warranted between the layout in fig. 4a, with variable parts *L* and *R*, and that AZee expression with the same variable parts. Others can involve more elaborate AZee constructions.

More layouts are then added for AZee coverage when no sufficient spontaneous regularity was observed in VD. This applies to the set of rules supporting the basic sign vocabulary (every sign needs an icon), but also the combining, structuring rules. For example, no stable VD layout was established for *instance-of*<sup>1</sup>, so we created a layout for it, shown in fig. 4b.

AZVD also allows to map a similar AZee output from multiple graphical layouts, as was observed in VDs. For example, straight separation bars corresponding to the meaning and form of AZee rule *in-context*<sup>2</sup> are commonplace in VDs (one horizontal instance is visible in fig. 2). They can be oriented in different ways, hence the definition of two *variants* of the same mapping (fig. 5).

Recursively then, any full AZVD combination determines a single AZee expression output. And since any AZee expression determines a single SL production as a result, AZVD guarantees that every diagram ultimately determines a single read-out, making it synthesisable in a testable manner.

For example, fig. 6 shows the AZVD for the full 2B-JP entry of the *40 brèves corpus*<sup>3</sup> (Filhol and Challant, 2022), whose signed production lasts 27 seconds and meaning is the following:

<sup>1</sup>Meaning of *instance-of* (*type*, *elt*): *elt*, understood as an instance of *type*.

<sup>2</sup>Meaning of *in-context* (*context*, *process*): event or state *process*, which happened in situation *context* or after *context* has happened.

<sup>3</sup>Each of the 120 entries consists in a video LSF translation for a French news item, and the AZee expression that represents it. <https://www.ortolang.fr/market/corpora/40-breves>

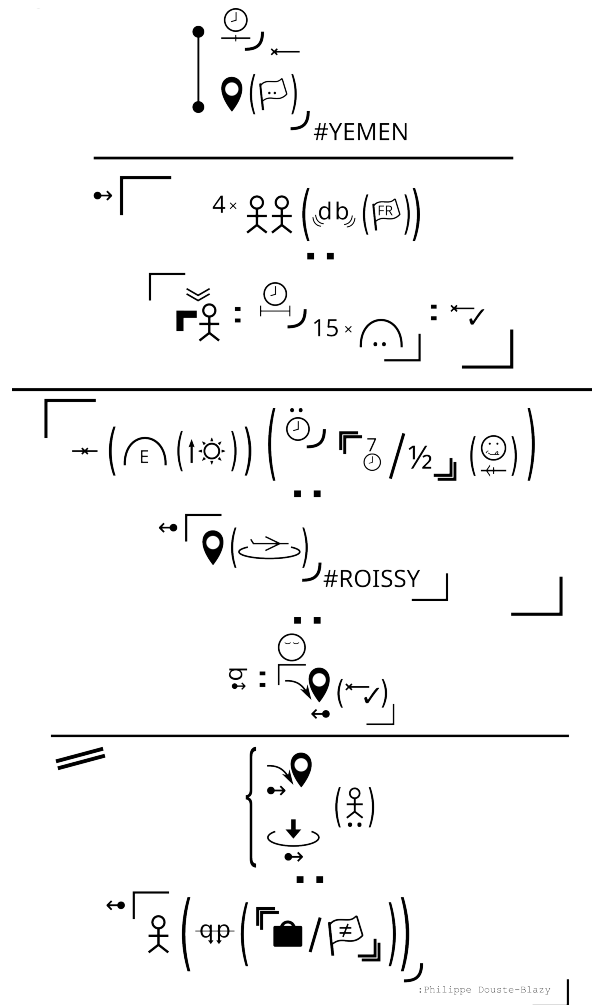


Figure 6: Example of AZVD

The four French tourists who were kidnapped 15 days in Yemen arrived on Wednesday, shortly before 7:30am, at Roissy airport, met by Minister of Foreign Affairs Philippe Douste-Blazy.

The AZVD in the figure exactly maps to the reference AZee expression, which in turn evaluates to a timeline specifying the necessary articulations for an avatar to render the same utterance.

AZVD is therefore a graphical system that is both synthesisable in principle and built with a method intended to maximise its adoptability. Testing synthesizability is verifying that an avatar animation can be rendered automatically from the expressions generated by composed graphical input, i.e. essentially AZee synthesis. To test adoptability, we ultimately need to place the system in the hands of users and involve the community in an iterative evaluation and improvement loop. To allow this process, the first step was to develop a software editor, able to assist in drawing AZVD in a controlled manner.





Figure 7: Screenshot of the AZVD editor: icon and layout menu on the left; main editor canvas in the middle; generated AZee output on the right

## 4. A software editor for AZVD

To enable testing of the AZVD proposition, we developed a software editor supporting the creation and manipulation of AZVD content. As we expect it to evolve with the AZVD system itself, we made the two following design choices:

- develop the editor as a web application to avoid requiring any installation or updating process on the user end, and enable instant deployment across all users on server upgrades;
- keep AZVD-side specifications separate from the server and load them dynamically on browser page load, in order to allow as much AZVD evolution as possible without changing the core application code.

After an overview of the chosen user interface, this section explains what the necessary AZVD components are, and how they are specified separately.

### 4.1. User interface

Inspired by the common WYSIWYG<sup>4</sup> interfaces to similar graphical content creation tasks, such as *Qt Designer* (windowed GUI design) or *Dia* (2D diagram drawing), we opted for a window layout with a central *canvas* to edit the AZVD content, and elements available in a left-hand *menu* to populate it with through drag-and-drop operations. We also added a right-hand *output panel* to display the generated AZee expressions, as we have stated the goal and benefit that every diagram determines one, and one only. This output synchronously reacts to every change on the canvas. A screenshot of the interface is given in fig. 7.

The top-level unit of AZVD specification is the left-hand menu object, which must contain information on both what to draw when inserted on the canvas and what AZee expression to generate as output. This is close to what has been called an “AZVD

<sup>4</sup>“what you see is what you get”

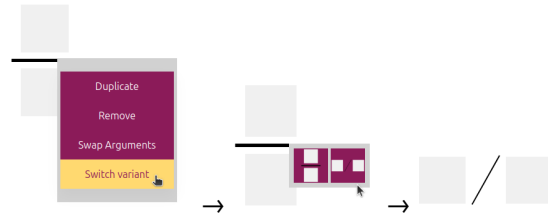


Figure 8: Switching from vertical to horizontal variant, specified in the same JSON spec file

mapping” up to here. The way to specify them for the editor is described in the next section.

### 4.2. Menu entries

As introduced earlier, adding, removing or changing AZVD mappings should be possible outside of the server implementation, whether to specify a graphical layout, icon or AZee output. At the moment this is done by providing JSON specification files, dynamically populating the menu on page load according to the specified content.

Each JSON file lists at least a description of a graphical layout (or fixed icon) and an AZee template to output when used on the canvas. To relate variants in the interface and pack them in a single menu entry, we allowed several layouts to be specified together in the same spec file. For example, both variants in fig. 5 can be specified together, in this case both mapping to the same parameterised AZee expression. This allows easy switching between variants of elements already on the canvas, as illustrated in fig. 8.

This explains how menu entries are created. The next two sections respectively deal with how to specify graphical layouts and the corresponding AZee output expressions.

### 4.3. Graphical layouts

In the general case, layouts are composed of one or more elements, each of which can be fixed graphics (e.g. an icon or line) or a variable part. Specification of a graphical layout is a problem of alignment and scaling of those contained elements. For example, the layout of fig. 5a is a group of three elements (two nested diagrams *context* and *process*, and a horizontal bar between them), aligned vertically through the centre, equally spaced, and the width of the middle bar constrained to be a little longer than the widest of the other two by a few points.

To do this, we first defined primitive element types to include in a layout:

- scalable graphics, rendered as specified directly inline with standard SVG code;
- text, which is rendered as a label verbatim in

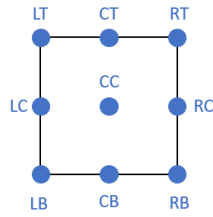


Figure 9: Generic element hotspots, named after horizontal and vertical positions relative to bounding box (L=left; T=top; C=centre; R=right; B=bottom)

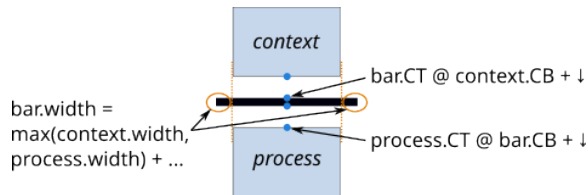


Figure 10: Specification of the layout in fig. 5a

the diagram, creating its own graphical bounding box;

- *drop zones*, which stand for the named variable parts of the layout, e.g. *context* and *process* above, to be filled for a complete diagram—they are rendered as plain grey boxes when empty.

Secondly, we implemented a relative positioning system based on generic hotspots assumed for any layout element, shown in fig. 9. Each new element in a layout is inserted by positioning one of its hotspots relatively to another's, with or without an offset expressed in absolute terms or relatively to other elements' sizes. Scaling or resizing elements is also possible, on either or both axes, proportionally or separately, in absolute terms or relatively to other elements' sizes.

Fig. 10 collects all the specifications required for the example layout of fig. 5a. It includes two "CT under CB" positioning constraints, which stack and centre the elements one under the previous, and one width scaling constraint on the horizontal bar, expressed as a function of the other elements' widths. This way, the width of the bar will adjust dynamically when the content of either drop zone is modified.

#### 4.4. AZee output

Every layout must specify the AZee expression it maps to, so that the AZee output panel be immediately updated with the new content when the layout is placed on the canvas. This can be a simple case of a fixed expression from a fixed layout, or one with variable parts like those in figures 4 and 5.

If the layout contains variable parts supported by drop zones, the output depends on their content, provided by further graphics filled in by the user. The AZee output then depends on the expressions that this content generates. In such case, the output specification can refer to the AZee for the nested content using a provided operator and the names of the zones, just like figures 4 and 5 used the same variable names as the labels in the corresponding layouts. Empty drop zones (incomplete diagrams) will generate a placeholder label to stand for the missing content, and dropping any graphical content in an empty drop zone will automatically update the output by expanding the placeholder to reflect the change.

## 5. Evaluation of current progress

The editor has reached a technically usable state, and we are gradually providing AZVD mappings for the AZee production set of LSF, as explained in section 3, to populate the menu. Straight away however, we note that graphical coverage of the entire production set is an unreasonable target to condition first tests on.

One reason is the size and open-endedness of the sign vocabulary. Looking at the *40 brèves* corpus alone, which serves as the AZee reference for LSF today (totals 1 hour of AZee-encoded LSF discourse), we find that out of the 858 distinct production rules applied, 768 are defined with no mandatory arguments<sup>5</sup>. Besides, we believe that users should be given priority to propose the icon graphics, debate choices<sup>6</sup> and possibly feed back to one another after some practice. Therefore, the effort to create enough individual icons to cover any significant portion of the vocabulary appears greater than we can afford without a dedicated team. It would also only serve as a kick-start proposition to be entirely reviewed anyway. But to provide enough vocabulary for the sake of demonstration, we decided to choose 5 entries of the corpus of which to cover the vocabulary entirely, namely 1A-OC, 1B-JP, 1O-VF, 1R-JP and 2B-JP. This represents a vocabulary set of 114 signs.

To insert signs—or indeed any signed piece of discourse—with no graphical solution yet, we created an alternative to AZVD mappings, namely

<sup>5</sup>This is to us the best characterisation of a vocabulary—or *lexical*—unit in AZee: a signed production that can be delivered without contextual input (a "citation", "canonical" form).

<sup>6</sup>For example, should it capture the meaning (promote a logographic symbol) or the articulated form (compose a phonographic encoding) of the represented sign? We have already documented the fact that spontaneous productions exhibit a logographic prevalence overall, but not an exclusive one (Filhol, 2020b).

AZee boxes. An AZee box can be dropped on the canvas instead of a regular graphical layout, and filled with AZee code, which will directly serve as its own mapped output. For a sign without an icon defined, an AZee box can therefore be used, filled with a simple named rule application, looking essentially like a gloss until an icon is defined. An example is visible at the bottom of fig. 6 (“:Philippe Douste-Blazy”), which is a name-sign for a prior member of the French government, for which we thought creating an icon was unnecessary.

Vocabulary signs aside, we are left with the production rules requiring at least an argument when applied. In our 40 brèves count, that remainder consists in 90 rules of the featured set:

- 12 types of pointing gestures (e.g. using index or hand sweep);
- 20 rules representing objects referred to as “classifiers” in the literature (e.g. `prf-flat-surface`, `prf-person-standing`);
- 58 recursive rules of various arities (unary rules like `with-worry`, binary ones like `info-about`, etc.).

The 58 recursive rules are the most interesting to cover as they are those building up the backbone structure of the discourse expressions. They typically have higher frequencies, and constitute a set that is much less open-ended than the sign vocabulary, in other words less subject to subsequent extensions. This is in a sense a more *grammatical* set, and securing mappings for it is a lot more stable an achievement than covering any vocabulary subset. Incidentally, and contrary to the lexical set known to be more significantly different between SLs, recent experiments seem to indicate that this set may be mostly transparent across different SLs (McDonald et al., 2024 (to be published)). It is something of interest if we later want to consider AZVD beyond its application to LSF.

We have covered all rules of that set with a working graphical layout and AZee mapping in the editor, except:

- 5 rules related to classifier use and geometric placement in signing space (`landmark-in-place`, `place-object`, `mult-around`, `mult-in-a-row`, `deploy-shape`);
- 4 rules supporting the logic for numbers above 20 (built with multipliers and sums) and doubled letters in fingerspelling—although these rules will not require graphics because numbers and words to fingerspell will appear spelt out in diagrams without being broken down (but an extension to the AZee output generation language from these text units will be necessary).



Figure 11: AZVD mapping for “pointage index (*target*)”

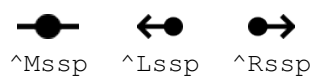


Figure 12: Basic AZVD point layouts

Let us now consider the pointing rules. Their signatures all resemble vocabulary signs, only they usually require a *target* argument of type POINT, and sometimes other geometric arguments, e.g. for orientation in a plane. Accounting for such a rule with AZVD is therefore comparable to finding an icon for a dictionary sign, only a non-optional variable part must be part of the layout. Fig. 11 shows the layout defined for `pointage index`, by far the most frequent: 256 occurrences in the corpus, over 4 times as many as the second-ranked, and indeed the 6th most frequent rule all together. Notice that it features a variable part, awaiting a point expression.

To enable filling such point arguments, we defined three more mappings, from the symbols shown in fig. 12 to the most basic and frequent point expressions in the corpus. These are `^Mssp` (neutral, central point of the signing space at about a forearm’s length of the signer’s abdomen), and `^Lssp` and `^Rssp` (points on either side of it, left and right respectively).

The more complex geometric point constructions or signing space references will require an AZee box at this point of our progress, and providing it the AZee code explicitly. The remaining 20 production rules in the above count, related to classifiers, have also not been accounted for yet. We come back to those in the prospects below.

In summary, aside from complex number and geometric constructions, we have reached most of the grammatical production set for LSF already. For instance, the 2B-JP entry of the 40 brèves corpus, shown in fig. 6, is fully editable within the program.

Fig. 13 illustrates a few steps of an AZVD construction, with the corresponding AZee output for each, for an LSF production meaning “French person returns from Portugal”, with Portugal located on the right-hand side of the signing space first, and the person returning to the left-hand side at the end. A recorded video of the whole process is available at <https://zenodo.org/records/10890951>. Note how every drop, move or swap action on the canvas updates the AZee output accordingly.

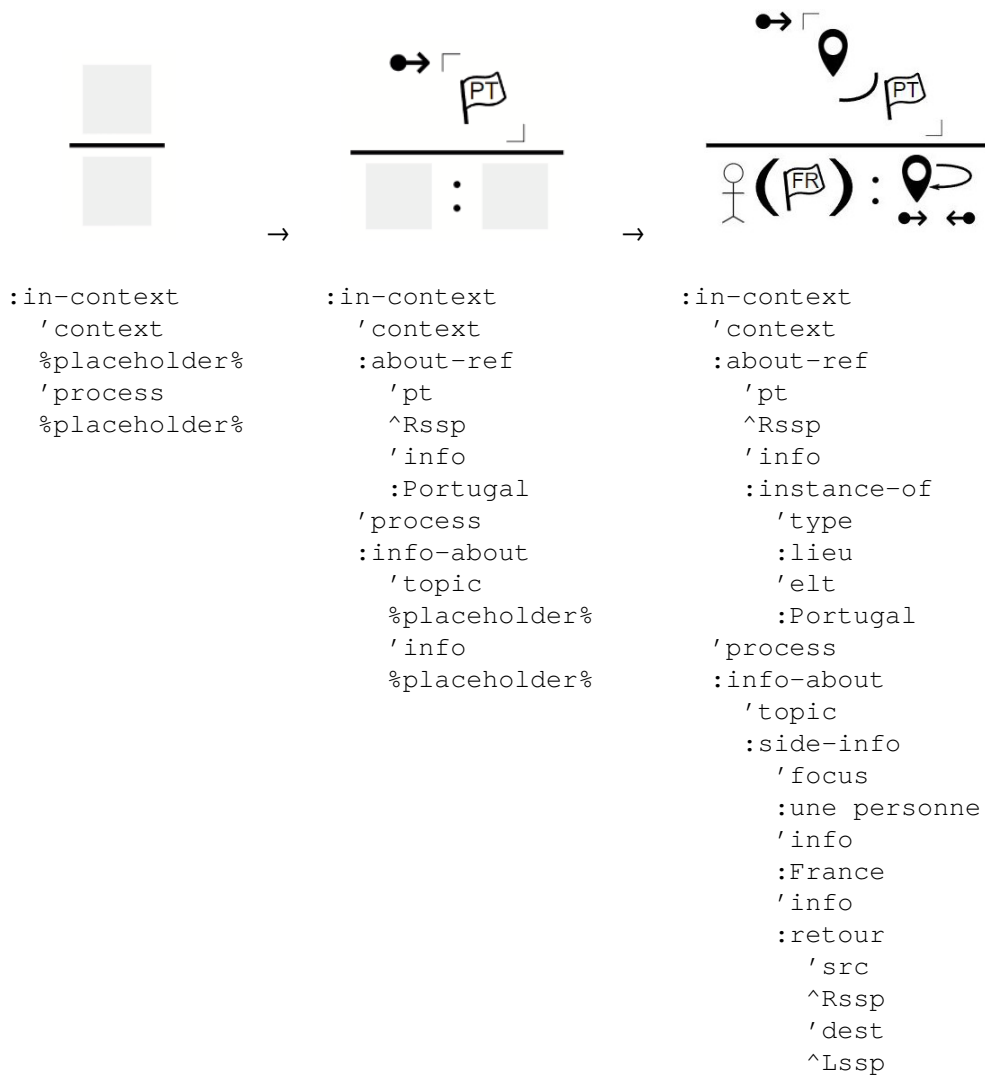


Figure 13: Progressive construction of an AZVD with the editor. NB in French: “lieu” = place, location; “une personne” = a person; “retour” = return.

## 6. Conclusion and future work

After reviewing a spontaneous practice of drawings to represent SL, we presented the AZVD system aiming to propose a graphical system both synthesisable and adoptable by SL users. We followed by presenting a software editor developed to support creation and editing of diagrams in the AZVD format.

The point of the editor in the long run is to allow users to apprehend the AZVD approach, evaluate its adoptability, and involve them in the system’s evolution as much as they would like to. But measuring adoptability with users through the editor can only be conducted reliably if it is fit to support AZVD manipulation transparently enough in the first place. An incomplete or non-ergonomic, counter-intuitive interface can indeed lead to rejection of AZVD as a whole even if the cause is the editor alone.

So to avoid this bias, we must separate the evaluation of the editor and that of AZVD as a scripting system. We will do so by taking a first step testing the application essentially as an AZee editor first. That is, measure how AZee experts feel assisted in the task of writing and reviewing AZee expressions. Any piece of AZee not covered with AZVD graphics can still be expressed in AZee code inside AZee boxes, which AZee coders would have done anyway without the editor. Reaching a positive evaluation on that aspect would constitute evidence that the interface and features of the editor provide enough comfort and assistance to allow users to direct their judgement at the manipulated script, not the manipulation tool.

Now even with a good editor, limitations to AZVD remain which should also be addressed. The most limiting factors are the missing layouts for the geometric rules and constructions (involving classifiers)



and the sign vocabulary (lexical set). Each of these two aspects represents a work prospect to increase the scope of AZVD graphics.

Geometric/classifier constructions were postponed mostly because a parallel work to encode the *Mocap1* corpus (LIMSI, 2020) with AZee expressions is in progress. This substantial work should result in a more stable reference for AZee representation of those constructions, which was to us an interesting contribution to wait for before defining a graphical layer for them. However, it is already clear that their infinite range in signed locations, paths, dynamics and classifier options does not come from an ever-growing set of ad hoc rules, but from the generative power and combinatorics of a limited set. We therefore believe tentative solutions should be in reach, similar to those addressing the grammatical set, only certainly requiring more layouts for native geometric objects (points, vectors, paths). This is to at least remove the constant need for AZee boxes in the diagrams, and propose a first graphical scheme to the discussion along with the other grammatical rules.

In contrast though, as explained above, it is impossible to do the same with the open-ended set of vocabulary signs. The prospect for us here, aside from keeping the possibility of glossing (a strategy well captured by AZee boxes already), is to allow to fallback on custom graphical choices, and ideally integrate a proposal and voting system, or existing lexically-oriented phonographic systems such as SignWriting or HamNoSys (fig. 1). Choices could be up- or downvoted by the community, and we would get to observe discrepancy or consensus in propositions. How variable are the logographic choices? How often do phonographic ones make spontaneous use of the existing systems? Much is yet to be learnt, on top of what VDs already exhibit, about how SL users envision scripting their language symbolically.

In the mean time, one already pictures the kind of diagrams AZVD allows to build, and notices two major differences with the prior systems. First, logography is allowed and frequently used in the graphics. We have already said that it played a major part in spontaneous VDs, while being totally absent in the other systems. Second, the diagrams exhibit the meaningful links between their constituents, reflecting the underlying structure of the utterances. This is very similar to the spontaneous VDs, which rarely present entirely separate parts, and rather keep them connected in a planar (2D) drawing. It also greatly contrasts with the other systems, which impose to follow the production sequence one lexical unit after the other, without connecting them in any meaningful way. If we trust the idea that following spontaneous practice is likely to favour adoptability, both of those properties are therefore

welcome.

Finally, we would like to leverage the fact that AZVD was designed not only to maximise adoptability, but also to be synthesisable. Integrating an avatar to the interface, for example under or instead of the AZee output panel, to render the AZVD canvas content would bridge over the full pipeline from AZVD editing to dynamic display of the scripted signed discourse. We are preparing for this exciting prospect, which in our view will even allow users with no knowledge of AZee to learn AZVD directly.

This way we hope to put the system in the hands of the Sign Language community with as few obstacles as possible to appreciating the AZVD system. More than evaluating a fixed state from a single field test, we will hopefully engage users on a continuous improvement process, and fuel the discussion about graphical Sign Language writing, which is an unresolved issue yet.

## 7. Acknowledgement

This work has been funded by the EASIER (Intelligent Automatic Sign Language Translation) European project. Funding's agreement Horizon 2020 no. 101016982.

## 8. Bibliographical References

- Anaïs Athané. 2015. La schématisation : un travail original de préparation à la traduction de textes vers la langue des signes française. *Double Sens*, 4.
- Camille Challant and Michael Filhol. 2022. A first corpus of azeed discourse expressions. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marseille, France.
- Michael Filhol. 2020a. Elicitation and corpus of spontaneous sign language discourse representation diagrams. In *Proceedings of the 9th workshop on the representation and processing of sign languages*.
- Michael Filhol. 2020b. A human-editable sign language representation inspired by spontaneous productions... and a writing system? *Sign Language Studies*, 21(1).
- Donald A. Grushkin. 2017. [Writing signed languages: What for? what form?](#) *American Annals of the Deaf*, 161(5):509–527. Gallaudet University Press.
- Thomas Hanke. 2004. Hamnosys—representing sign language data in language resources and

- language processing contexts. In *Proceedings of the workshop on the Representation and Processing of Sign Languages*, pages 1–6. European Language Resources Association (ELRA).
- Mihoko Kato. 2008. A study of notation and sign writing systems for the deaf. *Intercultural Communication Studies*, 17(4):97–114.
- John McDonald and Michael Filhol. 2021. [Natural Synthesis of Productive Forms from Structured Descriptions of Sign Language](#). *Machine Translation*.
- John McDonald, Rosalee Wolfe, Eleni Efthimiou, and Evita Fotinea. 2024 (to be published). Multilingual synthesis of depictions through structured descriptions of sign: An initial case study. In *Proceedings of the workshop on the Representation and Processing of Sign Languages*, Torino, Italy.
- Elena Antinoro Pizzuto and Paola Pietrandrea. 2001. [The notation of signed texts: Open questions and indications for further research](#). *Sign Language & Linguistics*, 4(1–2):29–45.
- S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. 1989. Hamnosys version 2.0, hamburg notation system for sign languages, an introductory guide. *International studies on Sign Language communication of the Deaf*, 5. Signum press, Hamburg.
- William C. Stokoe, Dorothy C. Casterline, and Carl G. Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Washington, DC.
- Valerie Sutton. 2014. *Lessons in SignWriting*, 4th edition. The SignWriting Press.

## 9. Language Resource References

- Filhol, Michael and Challant, Camille. 2022. [40 brèves](#). 2, ISLRN 988-557-796-786-3.
- LIMSI, CIAMS. 2020. [MOCAP1](#). ISLRN 502-958-837-267-9. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

# Content Questions in Sign Language From theory to language description via corpus, experiments, and fieldwork

<sup>1</sup>Robert Gavrilesco , <sup>2</sup>Carlo Geraci , <sup>3</sup>Johanna Mesch 

<sup>1,2</sup>Institut Jean-Nicod Département d'études cognitives  
ENS EHESS CNRS PSL Research University Paris

<sup>3</sup>Stockholm University

<sup>1,2</sup>24, Rue Lhmond, 75005 Paris,

<sup>3</sup>Universitetsvägen 10C, 10691 Stockholm

groom7778@gmail.com, carlo.geraci76@gmail.com, johanna.mesch@ling.su.se

## Abstract

The theory of language structure informs us about what we should expect when we want to investigate a certain construction. However, reality is often richer than what theories predict. In this study, we start from a theoretically informed set of hypotheses about the structure of wh-questions in sign language, we test them using a sign language corpus, a designed production experiment, and structured fieldwork in three sign languages, Swedish, Greek and French Sign Languages. The results will inform us on what type of contribution each research method can provide to reach accurate language descriptions.

**Keywords:** Sign Language Methodology, Content questions, Wh-sign

## 1. Introduction

The body of research on questions in sign language has been conducted either using typological questionnaires (Zeshan, 2006), fieldwork elicitation (i.a., Cecchetto et al., 2009; Neidle et al., 2000; Petro-rio and Lillo-Martin, 1997), or semi-formal experiments using pre-set elicitation materials (Geraci et al., 2015). To our knowledge, no corpus study has ever been conducted yet on the structure of content questions in sign language. In this work, we will use constituent questions as a case study to illustrate how a broad research question like the description of constituent questions in sign languages can be addressed using different methodologies, and the degree to which they yield comparable results. The purpose of this methodological exercise is not that of identifying the most appropriate method to study sign language syntax, but rather, to illustrate what a researcher can reasonably expect to find using one of the three traditional resources of language data, namely corpus, experiments, and fieldwork, which are treated here as case studies. In the remainder of the paper, we will present a brief overview of the relevant components of sign language content questions both from the perspective of the empirical description of the grammars of sign languages and from the perspective of the theoretical challenges that these constructions represent for formal approaches to language (Section 2). The methods for each case study are then described in Section 3, while in Section 4 the results are presented. In Section 5, we will offer a comparative discussion, while Section 6 concludes the paper.

## 2. Content questions in SL

Question formation is one of the most investigated topics in sign language syntax. This is due both to empirical and theoretical reasons. The empirical reason is relatively easy to imagine and has to do with the importance of describing main clause types, hence question description is often next to the description of declarative clauses, as opposed for instance to imperatives and exclamatives, which are much less investigated in sign language (Cecchetto, 2012). The theoretical reasons, however, are much more intriguing because they reveal two aspects that make sign languages different from spoken languages: one concerns the use of non-manual components as a distinctive marker for questions; the other concerns the position of wh-signs in content questions. The use of dedicated non-manual components, in particular facial expressions, to distinguish declaratives from questions has been described for both polar (yes-no) and content (wh-) questions. An example of non-manuals used in polar question is illustrated by the Italian Sign Language (LIS) examples in (1) below, where the declarative sentence and the polar question share the same sequence of signs, and are differentiated only by the non-manual components (see also Conte et al., 2010). Specifically, the head/torso is slightly forward and raised eyebrows spread throughout the sentence.<sup>1</sup>

<sup>1</sup>For a comprehensive study on polar questions in a sign language see Cañas (2021)

- (1) a. MUM MOVIES GO  
'Mum goes to the movies.'
- b.  $\overline{\text{MUM MOVIES GO}}^{\text{y/n}}$   
'Will mum go to the movies?'

As for non-manuals in content questions, furrowed eyebrows are very often described either to co-occur with the *wh*-sign only, or to spread over larger portions of the sentence. In American Sign Language (ASL), for instance, the *wh*-non-manual component spreads over the entire sentence if the *wh*-sign remains in argument position, while it can be limited to the *wh*-sign if it is found at the end of the sentence, as shown in (2) from Neidle et al. (2000).

- (2) a.  $\overline{\text{WHO LOVE JOHN}}^{\text{wh}}$   
'Who loves John?'
- b.  $\overline{\text{LOVE JOHN}}^{\text{wh}} \text{ WHO}$   
'Who loves John?'

The contribution of non-manual markers in questions is often compared to that of prosody in spoken language, because it can play a primary cue in sentence type detection as that, for instance, of rising intonation in languages like spoken Italian. For instance, polar questions are not syntactically differentiated from declaratives in spoken Italian (same word order, no question particles, etc.). They are, however, prosodically different because declaratives are typically associated with a falling intonation, while polar questions are normally associated with a rising intonation.

However, non-manuals have syntactic correlates that do not find an immediate equivalent in spoken languages. In fact, the distribution of interrogative non-manuals in ASL is associated with the *c*-command domain of the relevant projection (Neidle et al., 2000, but see Sandler, 2010 for a pure prosodic analysis), while it marks the syntactic chain in LIS (Cecchetto et al., 2009). This is best illustrated by the *in situ* content questions in (3). In fact, wide spreading crucially includes the subject (*c*-command domain) in ASL, while it is excluded in LIS.

- (3) a.  $\overline{\text{TEACHER LIPREAD}}^{\text{wh}} \text{ WHO}$   
'Who did the teacher lipread yesterday?'
- b.  $\overline{\text{PAOLO BOOK WHICH STEAL}}^{\text{wh}}$   
'Which book did Paolo steal?'

The second theoretical aspect concerns the fact that the privileged position for *wh*-signs in content questions often corresponds to the end of the sentence in several sign languages (Cecchetto, 2012).

Such clause final position, which is virtually unattested in spoken languages, is at the core of a debate in theoretical syntax since it seems in clear contrast with some of the basic tenets of contemporary syntax.<sup>2</sup>

### 3. Methodology

We took the sections about constituent questions of the SignGram blueprint as our starting point (Quer et al., 2017). As of today, the SignGram blueprint constitutes the most valuable resource for grammarians who are willing to begin a descriptive analysis of a sign language. Specifically, we focused on the *Syntax part, Chapter 1: Sentence type*. Section 2 of that chapter is devoted to interrogative sentences and it includes instructions on what to look for and provides references on how to elicit content questions. At the lexical level, the main topics to be covered are the identification of manual *wh*-signs and non-manual markers distinguishing content questions. At the sentential level, the main topics concern the distribution of *wh*-signs in the sentence, the scope of the non-manual markers, whether there are content questions without an overt *wh*-sign, the description of *wh*-phrases with a restriction (e.g., 'which student'), and whether it is possible to split the *wh*-sign from its restriction, the presence of *wh*-doubling, and multiple *wh*-questions.

We then looked into three sign languages, Greek Sign Language (GSL), French Sign Language (LSF), and Swedish Sign Language (STS), using a semi-formal production experiment, direct elicitation, and corpus resources, respectively. Ideally, these approaches replicate three real scenarios that a researcher might easily face with. We make them explicit here in the shape of case studies.

#### 3.1. Case Study 1: (Semi-formal) Production Experiment

A researcher decides to conduct a study on content questions in GSL. The language does not have an available corpus, and the department cannot hire a language consultant for that specific language. However, since the researcher is going to spend a couple of weeks in Athens, they decided to use their personal network of Greek Deaf friends, plus a mild snowball recruitment (Mouw et al., 2014) to conduct a semi-formal production experiment with the same stimuli used in Geraci et al. (2010,

<sup>2</sup>See for instance the debate about the position of *wh*-signs in ASL (Neidle et al., 2000; Petronio and Lillo-Martin, 1997) and the alternative analysis based on LIS data and tentatively extended to ASL (Cecchetto et al., 2009), while for the universal principles constraining the position of *ex-situ wh*-words see Kayne (1994).



2015), which have been reported to be a valuable resource by the blueprint.

The stimuli consisted of two pairs of pictures designed to mimic real-life situations like a car accident plus an insurance form (Fig. 1-2), and a domestic accident plus a medical form (Fig. 3-4). The task is assessed at pairs. One member of the pair receives a scene-picture, the other the corresponding form-picture. After they have looked at their picture, participants are asked to interact. Specifically, the person with the form picture is asked to fill in the form, playing either the role of a car insurance agent (Fig. 2), or the role of a doctor (Fig. 4). At the end of a trial, the participants change pictures and switch roles. These pictures have been designed specifically to elicit wh-questions in a semi-spontaneous environment. The participants are instructed not to follow the scenes strictly, but to take them as a hint to further elaborate the exchange. The forms, on the other hand, provide a memo for a wide variety of content questions (who, what, when, how, why, at what time, etc.).

Thirteen Deaf GSL signers participated to the study (7 pairs, one participant took part to two sessions to match a spare signer). The total duration of the recordings is of about 16 minutes. The dialogues are recorded with a phone camera and have been annotated using ELAN following the same template as in the corpus study (see below). The annotation (still on-going) is conducted by one of the author (Robert Gavrilescu), with the assistance of a GSL signer<sup>3</sup>.

### 3.2. Case Study 2: Elicitation study

Within a funded project to study some psycholinguistic aspects of the syntax of LSF, a researcher is asked to conduct a preliminary study on content questions. The study is necessary to provide essential information on how to properly construct the experimental stimuli. The LSF researcher does not have a large annotated corpus at their disposal, but can count on one/two language consultants who regularly collaborate with the linguistic group. They then decide to study content questions in LSF using the playback method (Schlenker, 2014; Lettieri et al., 2023). As illustrated in Lettieri et al. (2023), the playback method consists of a sequence of at least six steps:

- (4) a. Definition of the paradigm to investigate
- b. Recording the paradigm from one consultant
- c. Playing-back the paradigm to the informant(s)
- d. Recording acceptability and felicity judgments
- e. Discuss possible issues
- f. Repeat steps (4c-4e) at least once

<sup>3</sup>We are grateful to Dimitris Papapetrou for his help



Figure 1: Car accident: scene.

Figure 2: Car accident: form.

For this particular study, the scope of the research is given by the need of creating adequate stimuli for a psycholinguistic work, while the definition of the paradigm was given by the SignGram blueprint. The identification of wh-signs was done via LSF dictionaries and sign repositories (e.g., Spreadthesign Hilzensauer and Krammer, 2015 and Le Dico Elix). To illustrate how a subject wh-question paradigm was elicited, see the example in (5). The recording of the paradigm items was done by giving the language consultant a random sequence of signs (5a) to order in a grammatical sentence (5b) and then by substitution, asking to replace a noun with a wh-sign (5c), and reordering the signs in the sentence (5d-5e). Once one target sentence was finally reached, minimal variants are also recorded. Once the paradigm was obtained, in subsequent sections (at least a week apart) felicity and acceptability judgments were collected.

- (5) a. MOTHER, MARKET, SUNDAY, VEG., BUY  
Random sequence of signs
- b. SUNDAY POSS MOTHER BUY VEG. MARKET  
Baseline sentence  
'My mom bought vegetables at the market last Sunday.'
- c. SUNDAY **WHO** BUY VEG. MARKET



Figure 3: Home accident: scene.

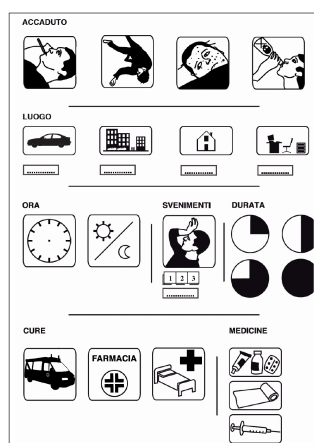


Figure 4: Home accident: form.

Target: wh-sign *in-situ* (substitution)

- d. **WHO SUNDAY BUY VEG. MARKET**  
Target: wh-sign in initial position (reordering)
- e. **SUNDAY BUY VEGETABLES MARKET WHO**  
Target: wh-sign in final position (reordering)  
'Who bought vegetables at the market last Sunday?'

The data for this study were recorded during 13 sessions, while judgments were collected during 4 sessions. Data from other projects were also collected within a session so that in a typical two-hour session, an alternation between tasks (recording and judgments) and projects (content questions, subordination, phonemic inventory, etc.) was guaranteed. This procedure avoids heavy and boring sessions on a single topic.

### 3.3. Case Study 3: Corpus study

Stockholm University has a large STS corpus which has been annotated since 2009 (or since 2003 if the ECHO project is included). The first release was in 2012, and a later release in 2021 contained the gloss tier fully annotated (Mesch, 2023; Börstell et al., 2016). The corpus contains free conversations, presentations, and elicited narrative tasks (e.g., the *Frog Story*), but nothing similar to the task used in the Case Study 2. Since no systematic description of content questions is available for the language, the researcher decides to look into the corpus and see what type of information is available. The corpus contains 190,000 tokens, from 42 participants from three regions of the country; and it has already been successfully used to study valency (Börstell et al., 2019) and the syntax-prosody interface (Puupponen et al., 2016).

The corpus search was done by looking both at wh-signs in the gloss tier and wh-words in the translation tier. A manual check was then used to exclude sentences in which wh-phrases are used in non-interrogative sentences (e.g., relative clauses). Since no systematic description of wh-questions is available for the language, new annotation tiers specific to the project have been added: question type, wh-position, position of nominal element in restricted wh-phrases, distribution of the non-manuals. These are intended to be used as potential dependent variables or categorical predictors in quantitative analyses with the levels indicated in (6):

- (6) a. **question type:** direct, embedded, constructed action
- b. **wh-position:** initial, final, in-situ, duplicated
- c. **Restricted wh-phrases:** adjacent to the wh-sign, split
- d. **distribution of non-manuals:** Absent, 1 sign, 2 signs, 3 signs, more

The annotation (still ongoing) is conducted by one of the authors (Johanna Mesch), who is also part of the research group that is responsible for the STS corpus at the University of Stockholm (see figure 5).

## 4. Preliminary Results

As for the inventory of wh-signs, all three methods of research have been able to spot a wide range wh-signs, indicating that the three languages have dedicated wh-forms for specified syntactic and semantic functions: **WHO** for animate/human individuals, **WHAT** for inanimate individuals in argument position, **WHERE** for locatives, etc. LSF combines specific

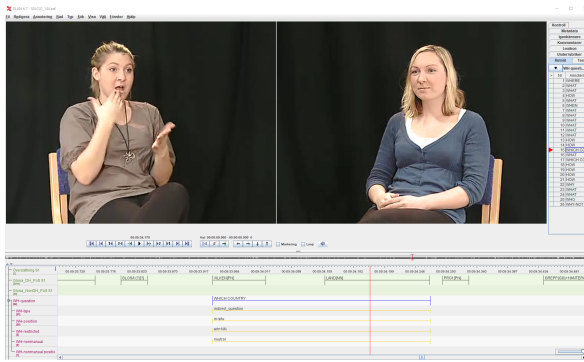


Figure 5: Corpus mining with coding schedule for the STS Corpus.

wh-signs depending on the restriction e.g., PRESIDENT WHO (*which president*), BOOK WHAT (*which book*), etc. STS uses the sign for WHO/WHICH in all types of restricted wh-phrases (the equivalent of English *which*), although there are cases in which the sign for WHAT is also used (e.g., WHAT REASON) No *which*-questions were found for GSL. One particular use of the sign for HOW was found in STS. The sign is used to create a sort of tag question eliciting an opinion from the addressee, as shown by the example in (7).

- (7) **Signer A:** STOP AGAIN YES OR HOW  
 'Stop, (do a recording) again, right?'  
**Signer B:** YES  
 'Yes.'

No variation among wh-signs is documented for LSF or STS, although it is known that there is a variant for the sign for WHO that is used in some regions of Sweden. Variation for the sign WHAT was found in GSL, where a two-handed palm-up sign (Fig. 6 right) or a two-handed 1-handshape form can be used (Fig. 6 left). The latter form is used by signers from the area of Athens.

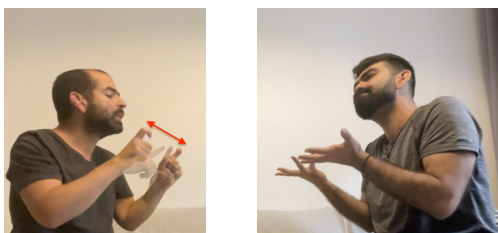


Figure 6: WHAT in GSL. Standard variant (right) and Athens variant (left).

Wh-questions without an overt wh-sign are documented in all three languages. Specifically, wh-phrases like *what time*, *how old*, and *how many* are often produced without a manual wh-sign (see Fig. 7, but are marked with the specific wh-non-manuals (see below).



Figure 7: HOW-MANY in STS. Only the sign MANY is produced.

Moving on to the syntactic part, the preliminary annotation of three videos of the production task returned 23 content questions in GSL, while approximately 250 content questions were recorded with the fieldwork method for LSF. The search for wh-signs in STS returned 2051 hits. Since, the STS corpus does not have an annotation tier for sentence type (declarative, interrogative, imperative, exclamative), a cross search to remove uses of wh-signs in non-interrogatives could not be performed at this stage. Nonetheless, a qualitative analyses of the corpus data is possible.

Indirect questions have been obtained for all languages. An example from GSL is given in (8).

- (8) ASK TIME ACCIDENT APPROX  
 'I am asking at what time approximately the accident happened.'

Content questions within a constructed action (role shift) are found in the STS corpus, while they have not been found in the production task, and were not elicited as part of the fieldwork activities. Two examples from STS are given in (9).

- (9) a. BOY      SEARCH      CALL      VOICE  
                     constructed action  
         WHERE FROG WHERE FROG  
         'The boy searches and calls for the frog.'
- b. MAN DS:PICK-UP WHO POSS IX-ON-GLASS  
                     constructed action  
         'When the window cleaner found the beer glass, he wondered whose it was.'

Concerning the position of wh-signs in the sentence, LSF allows wh-signs to remain *in situ*, to be found in sentence final position (after a locative phrase) and in sentence initial position (before a temporal adverb), as shown in (5) above. The fieldwork study revealed that the most preferred options are the *in situ* position (5c) and the sentence final position (5d), with the sentence initial position slightly marked.

For GSL, wh-signs are found in final position (10a), initial position (10a), and duplicated in initial and final position (10c).

- (10) a. IX2 COME HOW  
'How did you come (here)?'
- b. HOW CITY SAY  
'How do you say it was a city?'
- c. WHY COME WHY  
'Why did you come?'

For STS, wh-signs can appear sentence initially (11a), finally (11b), repeated at both edges of the clause (11c), and it can be omitted (11d).

- (11) a. [...] HOW WHAT DO IX2 TODAY  
[...] And what are you doing today?'
- b. FILM FESTIVAL THINK COMPARE ÖREBRO  
STOCKHOLM TWO DIFFERENT WHAT  
'Although I mean what is the difference  
between the film festivals in Örebro and  
Stockholm, what is the difference?'
- c. HOW TEACH LANGUAGE HOW  
'How does the teaching take place purely  
linguistically?'
- d. POSS2 FIRST WORK TO-BE SAAB IX2  
MALMÖ IX2  
'What was your first job? Was it at SAAB,  
in Malmö?'

Moving to restricted wh-questions, LSF allows the restriction to be stranded (12a) or pied-piped along with the wh-sign (12b). Interestingly, when the restriction is stranded, the sentence becomes ambiguous between a reading in which the wh-sign is interpreted as restricted by the subject or the object, as indicated in the possible translations for (12b). Crucially, (12a) cannot be interpreted as a stranded restricted wh-question on the object.

- (12) a. WHO DOG SCRATCH CAT  
'Which dog scratched the cat?'
- b. DOG SCRATCH CAT WHO  
'Which cat did the dog scratch?'  
'Which dog scratched the cat?'

Restricted wh-questions are rare in the production task, so no conclusions can be drawn for GSL.

As for STS, the search returned 62 hits of restricted wh-phrases with the order wh-sign + noun (WHICH YEAR, WHICH CITY, ETC.), while only 7 hits of sequences of noun + wh-sign, indicating a strong preference for the order in which the wh-sign precedes its restriction. Interestingly, STS does not seem to differentiate the wh-sign based on the animacy of the restriction. In fact, the sign for who is used across the board in restricted wh-questions. Restricted wh-questions in STS illustrate another interesting aspect of the syntax of content questions in SL, namely the possibility of having partial copy of the wh-phrase. The example in (13a) shows a

case in which the wh-sign is repeated, while the restriction is duplicated in (13b). Example (13c) shows a case in which the restriction and the wh-sign are repeated but the restriction is only partially repeated with the alternating pronoun (i.e., only the grammatical features of the restriction are repeated, and not its encyclopedic content).

- (13) a. WHICH BOOK WHICH  
'Which book?'
- b. BRING BOOK WHICH NEW BOOK  
'Which new book did you bring?'
- c. TERRACED HOUSE WHICH EASY CONTACT  
NEIGHBOURS WHICH IX-alt  
'In which house was it easier to contact  
neighbours?'

Finally, turning to the non-manual components. These are present in all languages. As for GSL, the proper distribution is yet to be determined, but there seems to be a head leaning forward and a slight eyebrow raising in correspondence of the wh-signs, although this seems to be optional. As for LSF, the non-manuals attested in the sample are furrowed eyebrows and squinted eyes. They often co-occur with manual wh-signs, but there are tokens in which those non-manuals are absent. When they occur, they may spread over portions of the sentence larger than the wh-constituent, although this is not the most common option. STS non-manuals for wh-questions are similar to those of LSF (see Fig. 7), but they appear to have a larger spreading in the sense that the non-manuals co-occur with several signs and are not restricted to the wh-sign only.

## 5. Discussion

Although preliminary, the results reported in Section 4 reveal interesting aspects of each methodology.

The production task is particularly effective in eliciting short wh-questions, typical of the spontaneous interaction, as already documented for LIS (Geraci et al., 2015). Despite the small number of tokens, it also shows a considerable amount of syntactic variation illustrating that GSL allows wh-signs to occur at either edge of a clause and even repeated at both edges. Although the population sample was not selected for this purpose, the method is also robust enough to record some lexical variation and elicit complex constructions like embedded wh-questions. For different reasons, the particular task does not seem to be adequate to study questions inside constructed actions, or *in situ* wh-questions. In fact, the participants' roles in the task somehow prevent constructed actions from occurring. As for *in situ* wh-signs, considering the overall small number of signs per sentence, it is complicated to find



syntactic evidence of the correct position of the *wh*-sign in the sentence.

The corpus study provided a considerable number of hits, although some of them may not be genuine content questions. Since the corpus contains data from a variety of tasks (narratives, presentations, conversations), it is crucial to notice that most of the hits come from the conversation task (see figure 8). So, if one were to start a corpus annotation for a study on content questions, the advice is to start looking into conversation videos before narratives or presentations. At the syntactic

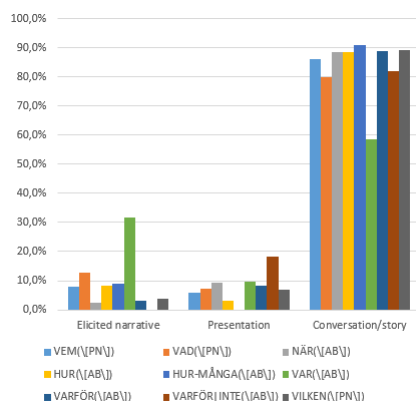


Figure 8: The distribution of *wh*-signs in the STS Corpus.

level, we could not evaluate the quantitative distribution of *wh*-signs because the annotation has not yet been completed. However, from a qualitative inspection, the data seem to be rich enough to determine the amount of variation in the position of *wh*-signs. The richness of the data will also allow an understanding of the distribution of restricted *wh*-questions. The corpus data also revealed the presence of questions inside constructed actions and tag constructions, which did not emerge from the production experiment and can be very hard to discover from fieldwork sessions.

Unfortunately, pure production data cannot provide negative evidence, this is true for both the experimental method and the corpus method. Specifically, understanding the conditions in which tag questions are acceptable might require the construction of *ad hoc* paradigms that might be better investigated using a different method.

As for fieldwork data, the identification of the target paradigms is much simpler to obtain than other with other methods and the possibility of getting negative evidence is something that is extremely valuable to create grammatical theories. At the level of grammatical description, fieldwork methods provide quick access to basic facts, but they are less suitable for capturing a wide range of variation. The method is ideal for a deep understanding of complex grammatical constructions (especially

with long sentences) but a bit less for pure exploration (and accidental discoveries). For instance, tag questions in STS would be very hard to discover using the elicited method, unless the researcher is already prompted about the existence of that construction and of what type of lexical material is needed.

Table 1 summarizes how the description of content questions can be accomplished using fieldwork, corpus, and experimental resources.

Level	Exper.	Fieldwk.	Corpus
Manual signs	ok	ok	ok
Non-manuals	*	ok	ok
Position of <i>wh</i> -signs	?	ok	ok
Content Q with no <i>wh</i> -sign	ok	ok	ok
Restricted <i>wh</i> -phrases	NA	ok	ok

Table 1: Summary of the descriptive adequacy of the three methods. ok = objective reached, \* = objective not reached, ? = objective partially reached, NA = not assessable.

Although these are only preliminary, the picture that emerges is that fieldwork and corpus methods provide similar results, proving adequate tools for linguistic description. On the other hand, the experimental task does not allow for a satisfactory analysis of the non-manuals and restricted *wh*-phrases, while the distribution of *wh*-signs in the sentence is only partially accomplished. We believe that this is due to the fact that the experimental task elicited very short questions. Short sentences are not ideal to analyze the spreading of non-manuals or the syntactic distribution of *wh*-signs because sentences with few signs do not allow to conclusively understand the underlying structure of the construction. Furthermore, the specific task was not designed to elicit restricted *wh*-questions. So, it is not a surprise that with the small sample we considered here none was actually produced. One final note on this methodology. Experimental studies are an excellent tool for hypothesis testing but are rarely used for descriptive purposes. However, if one were aiming to obtain a satisfactory description of the content question, more than one experiment is likely needed.

## 6. Conclusions

In this work, we addressed the methodological question of what types of information can be obtained when different methodologies are used to accomplish a similar task. We used three different case studies to explore how experimental, field-

work and corpus methods gather linguistic data to describe content questions in sign language. Overall, the results of the first case study, experiment data, offer a pilot of what can be further and more extensively explored with more controlled settings and more participants. Still, if this method is to be pursued, it should be paired with a comprehension study, although admittedly the analysis of complex constructions might reveal difficult using this method. The results of the second case study, elicited data, is a deep description of some aspects of the syntax of content questions in LSF with little exploration of variation and of the effects that variation may have on the constructions. In this respect, an experimental or a corpus study, if the resource is available, would be an ideal complement. The results of the third case study, corpus data, is a rich set of *wh*-constructions, which has only been qualitatively investigated, but that provided an interesting glance at the amount of variation in the language. The downside of this method, as already observed, is the lack of negative evidence, and the difficulty of probing the deep properties of the constructions. Hence, if a researcher starts with a corpus study, after a qualitative and quantitative analysis of the data, complementary fieldwork data are ideal.

## 7. Acknowledgements

Part of this research received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788077, Orisem, PI: Schlenker). Research was conducted at Institut d'Etudes Cognitives, École Normale Supérieure – PSL Research University. Institut d'Etudes Cognitives is supported by grants ANR-10-IDEX-0001-02 and FrontCog ANR-17-EURE-0017.

## 8. Bibliographical References

Carl Börstell, Tommi Jantunen, Vadim Kimmelman, Vanja De Lint, Johanna Mesch, and Marloes Oomen. 2019. [Transitivity prominence within and across modalities](#). *Open Linguistics*, 5(1):666–689.

Carl Börstell, Mats Wiren, Johanna Mesch, and Moa Gärdenfors. 2016. [Towards an annotation of syntactic structure in the Swedish Sign Language corpus](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 19–24, Portorož, Slovenia. European Language Resources Association (ELRA).

Sara Cañas. 2021. [Polar interrogatives in Catalan Sign Language \(LSC\): a comprehensive grammatical analysis](#). Ph.D. thesis, Universitat Pompeu Fabra.

Carlo Cecchetto. 2012. Sentence types. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign language. An international handbook*, pages 292–315. Mouton de Gruyter, Berlin.

Carlo Cecchetto, Carlo Geraci, and Sandro Zucchi. 2009. Another way to mark syntactic dependencies The case for right peripheral specifiers in sign languages. *Language*, 85(2):278–320.

Genny Conte, Mirko Santoro, Carlo Geraci, and Anna Cardinaletti. 2010. [Why are you raising your eyebrows?](#) In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 53–56, Valletta, Malta. European Language Resources Association (ELRA).

Carlo Geraci, Robert Bayley, Chiara Branchini, Anna Cardinaletti, Carlo Cecchetto, Caterina Donati, Serena Giudice, Emiliano Mereghetti, Fabio Poletti, Mirko Santoro, and Sandro Zucchi. 2010. [Building a corpus for Italian Sign Language. methodological issues and some preliminary results](#). In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 98–101, Valletta, Malta. European Language Resources Association (ELRA).

Carlo Geraci, Robert Bayley, Anna Cardinaletti, Carlo Cecchetto, and Caterina Donati. 2015. [Variation in Italian Sign Language \(LIS\): The case of \*wh\*-signs](#). *Linguistics*, 53(1):125–151.

Marlene Hilzensauer and Klaudia Krammer. 2015. [A multilingual dictionary for sign languages: Spreadthesign](#). Technical report, Alpen-Adria-Universität Klagenfurt, Klagenfurt.

Richard S. Kayne. 1994. *The antisymmetry of syntax*. MIT Press, Cambridge.

Jessica Lettieri, Mirko Santoro, and Carlo Geraci. 2023. [On Elicited Data in Sign Language Syntax. FEAST. Formal and Experimental Advances in Sign language Theory](#), 5:88–99.

Johanna Mesch. 2023. [Creating a multifaceted corpus of Swedish Sign Language](#). In Ella Wehrmeyer, editor, *Advances in Sign Language Corpus Linguistics*, chapter 9, pages 242–261. John Benjamins Publishing Company, Amsterdam.

- Ted Mouw, Sergio Chavez, Heather Edelblute, and Ashton Verdery. 2014. [Binational social networks and assimilation: A test of the importance of transnationalism](#). *Social Problems*, 61(3):329–359.
- Carol Neidle, Judy A. Kegl, Dawn Maclaughlin, Benjamin Bahan, and Robert G. Lee. 2000. *The Syntax of American Sign Language*. MIT Press, Cambridge, MA.
- Karen Petronio and Diane Lillo-Martin. 1997. Wh-Movement and the Position of Spec-CP: Evidence from American Sign Language. *Language*, 73(1):18–57.
- Anna Puupponen, Tommi Jantunen, and Johanna Mesch. 2016. [The alignment of head nods with syntactic units in Finnish sign language and Swedish sign language](#). *Proceedings of the International Conference on Speech Prosody*, pages 168–172.
- Josep Quer, Carlo Cecchetto, Caterina Donati, Carlo Geraci, Meltem Kelepir, Roland Pfau, and Markus Steinbach, editors. 2017. *SignGram Blueprint*. Mouton de Gruyter, Berlin.
- Wendy Sandler. 2010. [Prosody and syntax in sign languages](#). *Transactions of the Philological Society*, 108(3):298–328.
- Philippe Schlenker. 2014. [Iconic features](#). *Natural Language Semantics*, 22(4):299–356.
- Ulrike Zeshan, editor. 2006. *Interrogative and Negative Constructions in Sign Language*. Ishara Press, Nijmegen.

# Matignon-LSF: a Large Corpus of Interpreted French Sign Language

Julie Halbout\*<sup>1</sup>, Diandra Fabre\*<sup>2</sup>, Yanis Ouakrim\*<sup>1,2</sup>, Julie Lascar\*<sup>1</sup>  
Annelies Braffort<sup>1</sup>, Michèle Gouiffès<sup>1</sup>, Denis Beautemps<sup>2</sup>

<sup>1</sup>Univ. Paris-Saclay, CNRS, LISN, Orsay, France

<sup>2</sup>Univ. Grenoble Alpes, CNRS, GIPSA-Lab, Grenoble, France

<sup>1</sup>firstname.lastname@lisn.upsaclay.fr,

<sup>2</sup>firstname.lastname@gipsa-lab.grenoble-inp.fr

## Abstract

In this paper we present Matignon-LSF, the first dataset of interpreted French Sign Language (LSF) and one of the largest LSF dataset available for research to date. This is a dataset of live interpreted LSF during public speeches by the French government. The dataset comprises 39 hours of LSF videos with French language audio and corresponding subtitles. In addition to this data, we offer pre-computed video features (I3D). We provide a detailed analysis of the proposed dataset as well as some experimental results to demonstrate the interest of this novel dataset.

**Keywords:** French Sign Language, LSF, dataset, interpretation, alignment

## 1. Introduction

Automatic processing of sign languages (SL) is an expanding field, but unfortunately the vast majority of these languages are still poorly endowed in terms of corpora available for research. This is particularly the case for French Sign Language (LSF). One potential source of SL data is television (Koller et al., 2015; Albanie et al., 2021), where the number of interpreted programs has increased in recent years. However, the access to this data is generally not easy for research purposes, due to rights or technical problems. In France, weekly Council of Ministers debriefings yield [open-access](#) videos which are systematically interpreted in LSF. We have taken advantage of this opportunity to compile a new dataset called Matignon-LSF<sup>1</sup> (fig. 1), which is presented in this paper.

The primary language modality of the TV programs is speech. Speech may be subtitled, sometimes in real time, either automatically with all the potential errors that this entails, or in a live subtitling studio with time and format constraints. Speech may also be interpreted in SL, sometimes in post-production, which enables the SL version to be prepared and corrected, or sometimes in real time. In this last scenario, several phenomena occur. Usual practice in interpreting is for the professional to interpret into their native language. The situation is different in the case of SL interpreting because it is necessary for the interpreter to hear speech. Therefore, unless the interpreter is a CODA (child of deaf adult), he/she interprets into a second language. In addition, there is some evidence of differences between the output of hearing and deaf interpreters

\*These authors contributed equally to this work and none of the authors are Deaf

<sup>1</sup>Matignon refers to the official residence of the French Prime Minister, and in extends to the french government.



Figure 1: Screenshot from a video in the Matignon-LSF dataset, showing debriefings from the French government's Council of Ministers.

(Stone and Russell, 2011). Furthermore, due to strong time constraints, SL during real-time interpretation tends to closely follow the grammatical structure of the spoken language, with evidences that differences in forms of language are reduced in interpreted content (Dayter, 2019). The interpreters may choose not to convey information from the audio stream that they consider to be redundant to the visual stream of the footage. Fluent signers can generally tell the difference between interpreted and non-interpreted SL, as well as signing by native deaf signers and non-native or non-deaf signers.

It is worth emphasizing that, due to the interpretation process, the source language can interfere in the signing. Thus interpreted SL can be different from original SL (i.e. directly produced by signers). However, there is little work on describing or quantifying these differences.

Having said that, this kind of dataset may be very



useful in automatic processing because it provides more SL data, even if it is task specific. In our case, it also has the advantage of being open-data.

In this paper, after a brief overview of the corpora currently available (section 2), particularly in LSF, section 3 presents the Matignon-LSF dataset, the collection and processing of the data, and section 5 discusses the perspectives opened by this new dataset.

## 2. Related Work

As part of the recent Easier European project, an overview of existing datasets for the European SLs was drawn up (Kopf et al., 2023). These datasets were divided in two categories: linguistic corpora and broadcast data. The former offer high-quality data with rich transcriptions and annotations, while the latter are available in large quantities. Since the publication of this report, other datasets have been released, such as BSL-1K (Albanie et al., 2020) and more recently BOBSL in British Sign Language (BSL) (Albanie et al., 2021), which represents a change of scale in terms of dataset, providing researchers with over 1,200 hours of sign language interpreted from BBC broadcasts. In a similar vein, the American Sign Language YouTube-ASL dataset (Uthus et al., 2024) totals almost 1,000 hours of videos from the web. Also in ASL, the How2Sign corpus, published in 2023, is of particular interest, as it is the largest laboratory corpus of original (non interpreted or translated) SL. This has already been the subject of several works (Duarte et al., 2021).

LSF has been the subject of several corpora collections over the last 10 years (Braffort, 2022). Most of these LSF corpora have been compiled in laboratories mainly for linguistic research works, and have two main shortcomings: fully annotated datasets like *Rosetta* and *40 brèves* are very small, containing less than 4 hours of data and larger datasets, such as *Creagest* (Balvet et al., 2010), are only partially annotated. The *DictaSign* dataset, consisting of 8-hour dialogues (Belissen et al., 2020), is currently partially annotated. Nevertheless, it remains valuable for recognizing signs in context, including lexical (Ouakrim et al., 2023) and non-lexical instances (Belissen et al., 2020).

Recently, two LSF datasets have been made available to overcome these problems: *Mediapi-Skel* (Bull et al., 2019) and *Mediapi-RGB* (Ouakrim et al., 2024). The last one comprises 86 hours of videos in LSF produced by deaf reporters or presenters from the bilingual online medium *Média'Pi!* with French subtitles produced by deaf translators. The subtitles are well-aligned with LSF videos, and the dataset has been prepared for processing (Ouakrim et al., 2024). These two corpora are in a LSF-to-

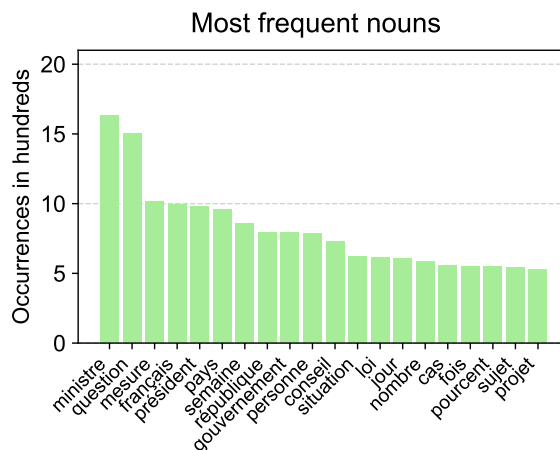


Figure 2: 20 most frequent nouns in the subtitles of Matignon-LSF.

French modality because subtitles were produced accordingly to the signing (and not the other way around) and are perfectly aligned. These two corpora are much larger than the previous ones in LSF, except for the non completely annotated *Creagest* corpus.

Due to the economic model of this medium, videos are unavailable for *Mediapi-Skel* and only partially available for *Mediapi-RGB*, which may be a limitation for researchers wishing to use features other than those pre-extracted by the authors (body pose, I3D, etc.).

Thus, our aim is to collect a new LSF dataset that is both large and open. We are therefore interested in interpretation data from French broadcast and created the *Matignon-LSF* dataset detailed in the following sections.

## 3. Dataset overview

French government's [Council of Ministers debriefings](#) take place once a week at l'Elysée. They are filmed, subtitled and, since July 2020, interpreted in LSF. The *Matignon-LSF* dataset is based on the LSF interpretations and subtitles of these debriefings. We do not have further information yet regarding the work process of the interpreters, but they probably don't have much material to prepare their interpretation. To date, it includes 67 debriefing videos. Figure 2 shows the 20 most frequent nouns of the dataset, demonstrating that the content of the speech is strongly related to French politics (top five words: *minister*, *question*, *measure*, *french* and *president*).

59 videos consist of the government spokesperson's speech (which varies from 4 minutes to 20 minutes, with an average of about 12 minutes), followed by a question-and-answer session with journalists. This part can vary depending on the

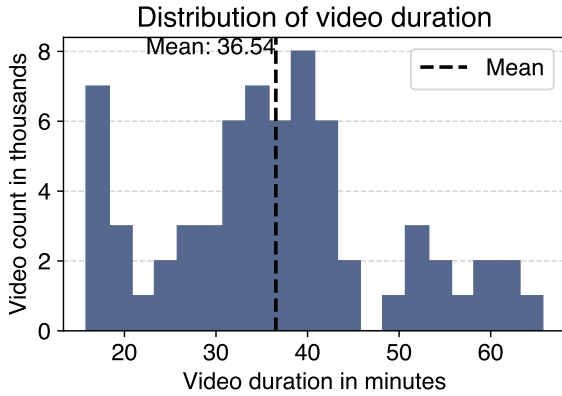


Figure 3: Video’s duration distribution.

topics and the number of journalists in the press room (from 8 minutes to almost an hour, with an average of 23 minutes). In five other videos, ministers are invited to present their points after the spokesperson’s speech, and they are asked questions in addition to the spokesperson. In the three remaining videos, the press conference is held without a spokesperson, and the ministers deliver their speeches directly, with a shorter question-and-answer session. The 67 delivered videos have a total duration of 39 hours, with an average duration of 36 minutes. The distribution of video duration is shown in Figure 3.

The subtitles (written French) in the dataset is composed of a total of 447k tokens for a total vocabulary size of  $10k^2$ . From the subtitles, we extracted 18k sentences, as described in section 4.3. Matignon-LSF features 15 signers.

The characteristics of the dataset are summarized in the table 1.

<b>Total duration (h)</b>	39
<b>#videos</b>	67
<b>#subtitles</b>	51131
<b>#sentences</b>	18000
<b>#french words vocab.</b>	10000
<b>#signers</b>	15
<b>#speakers</b>	3*
<b>Video resolution (px)</b>	$494 \times 494$
<b>Frame rate (fps)</b>	30

Table 1: **Dataset overview.** \*journalists and ministers not included.

To date, corpus Matignon-LSF lies between Mediapi-Skel and Mediapi-RGB in terms of size (fig. 4).

<sup>2</sup>we used SpaCy tokenizer <https://spacy.io/>

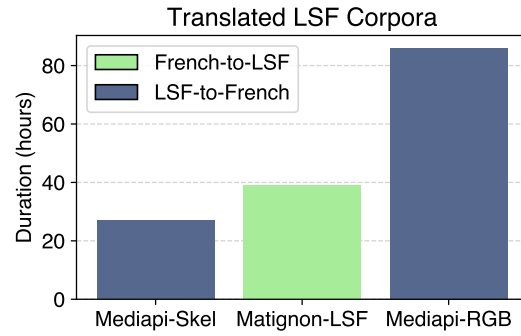


Figure 4: Duration of translated LSF Corpora. Matignon-LSF is the second largest translated LSF Corpora after Mediapi-RGB and the largest interpreted LSF corpora.

## 4. Data collection and processing

This section details the construction of the Matignon-LSF dataset. We present the raw data and the processing carried out to provide the dataset. The diverse processes are documented in a [GitHub repository](#), organized as a toolbox to enable reproduction and expansion of the corpus, as new press release takes place once a week.

### 4.1. Collecting the SL videos and subtitles

Each week, the debriefing is filmed and uploaded on [Youtube](#) and/or [Dailymotion](#) and comes with a corresponding set of written French subtitles aligned with the audio. Original videos have a resolution of 1080 px and a frame rate of 30 fps.

Using the [PyTube](#) Python library, we downloaded all videos issued between December 2020 and December 2023 along with their associated audio track. We then used the [YouTube Transcript Python Api](#) to download the subtitles, and keep only manually written subtitles, setting aside videos that only have generated subtitles. Obtained JSON files are then converted to the VTT subtitle format. Next, using [OpenCV](#), we crop the videos so as to retain only the square containing the LSF interpreter.

After the above steps, we obtain  $494 \times 494$  px LSF videos with associated French audio and subtitles.

### 4.2. Processing the videos

Skeleton keypoints, such as those provided by [OpenPose](#) (Cao et al., 2018) and [Mediapipe Holistic](#) (Lugaresi et al., 2019), are essential inputs for various automated sign language processing tasks. These tasks include cropping of hands or faces (Huang et al., 1994), generating sign language (Ventura et al., 2020), and improving recognition methods (Belissen et al., 2020).

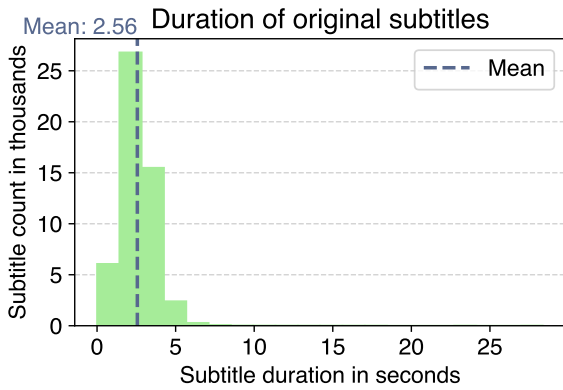


Figure 5: Distribution of subtitle duration before sentences extraction.

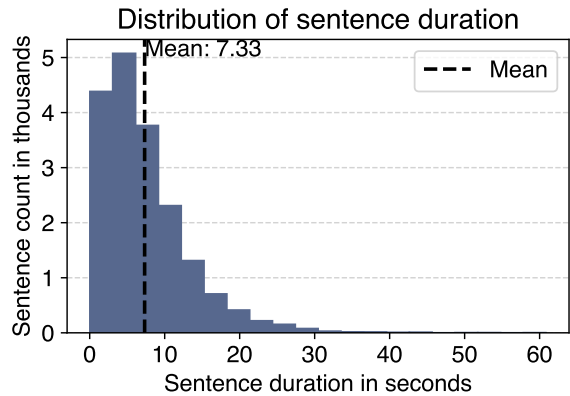


Figure 6: Distribution of sentence duration after sentences extraction.

Other automatic sign language processing methods (Tarrés et al., 2023; Renz et al., 2021) rely on features extracted from sign language videos by the I3D model (Carreira and Zisserman, 2017). We used this architecture to extract features from our videos. Specifically, we have used the fine-tuned model provided by Varol et al. (2021).

### 4.3. Processing the subtitles

As subtitles are constrained by length for display reason, they do not necessarily form sentences. However, the translation tasks often operate at the sentence level.

To address this, we generate a sentence-level segmentation from the subtitles. We adopt the same approach as Albanie et al. (2021) to build our sentence-segmented subtitle files. We split subtitles on sentence boundary punctuation. When a sentence spans multiple subtitles, it is easy to extract the sentence by concatenation. It is more complicated when multiple sentences fall in one subtitle. As the method used by Albanie et al. (2021), to preserve the alignment, we calculate the duration of a character (based on the subtitle’s characters length). We can use this information to associate a duration to each sentence within the subtitle. Then, we can calculate the new subtitle’s timestamps on this basis. The disparity of the subtitle’s duration between the original subtitles and the sentence-segmented subtitles is illustrated in Figures 5 and 6. The average time thus increases from 2.56 to 7.33 seconds.

The corpus will be soon deposited on the Ortolang platform and will be regularly updated over time. We estimate that it should be able to increase by around 13h per year.

## 5. Perspectives

The Matignon-LSF corpus has a number of advantages that can be exploited to address various computer vision and natural language processing tasks.

**Alignment.** At this stage, the French subtitles and LSF of Matignon-LSF are not yet aligned as can be seen in Fig 7. This example shows two consecutive sentences. “Un cap pour contrôler l’épidémie. Un cap pour relancer notre pays.” (*A direction to control the epidemic. A direction to relaunch our country.*). We observed that the length of the two signed sentences (4.64 seconds) is longer than that of the two spoken sentences (3.9 seconds). Therefore, a manual shift of the speech subtitles is not enough to fit the data: the GT and Sub alignments would start at the same time, but end differently.

Whatever the type of language (spoken, written or signed), machine translation methods require prior alignment between the source and the target languages. In order to use this dataset for translation tasks, it is necessary to be able to associate an extract of LSF with its corresponding French subtitles. The Matignon-LSF dataset contains a complete translation for each of the 67 videos. However, providing 35-minute video sequences ( $\pm 52,500$  frames) and their associated translations to a translation model would be very costly. It would therefore be necessary to divide these videos into sub-extracts.

State-of-the-art methods mostly rely on sentence segmentation. Hence, videos and text are split into sentence-like units, with an association between text and SL: for each SL sentence, the text corresponding to the translation is given. However, producing such an SL sentence/text alignment from an interpreted SL dataset is a real challenge: the text is aligned with the audio, whereas SL interpretation is performed with a latency that varies

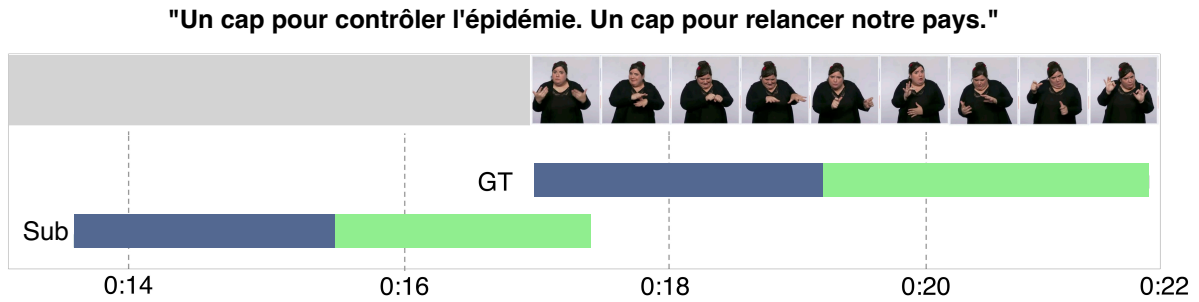


Figure 7: Demonstrating the alignment challenge in Matignon-LSF. The GT line corresponds to a manual alignment (or Ground Truth) annotated for this specific figure while the Sub line corresponds to the subtitles as provided with Matignon-LSF. Blue block corresponds to one sentence while green block corresponds to another sentence.

in time and from one interpreter to another. Thus the very first task to be carried out on this dataset should be to align the subtitles with the LSF content. Manual alignment requires a considerable time commitment as explained in (Bull et al., 2020): It takes an expert fluent in sign language approximately 10-15 hours to synchronize subtitles with one hour of continuous sign language video. Automatic alignment methods as the one used for the BOBSL dataset (Bull et al., 2021) could be a solution but might need some fine-tuning for LSF.

**Sign Language modeling.** The Matignon-LSF dataset can be used as it is, with no need for prior alignment, for sign language modeling and can be used to train unsupervised language models on LSF such as SignBERT (Hu et al., 2021).

**Sign Recognition.** With the help of a method like Lascar et al. (2024)’s automatic annotation process currently under development, we could perform automatic sign recognition and classification. This would provide information on the number of lexical signs in our dataset. Sign classification is also a step towards aligning our dataset between SL and the subtitles. However, one should note that the sign interpreters produce an interpretation of the speech that appears in the subtitles, as opposed to a transcription. This means that words in the subtitles may not correspond directly to individual signs produced by the interpreters, and vice versa. There may also be discrepancies between the audio and the subtitle text.

**Sign Language Translation.** Once aligned, the Matignon-LSF dataset could be used to train machine translation models for a wide variety of modalities: LSF to French text, LSF to Speech, and vice-versa (Ventura et al., 2020; Müller et al., 2023; Ouakrim et al., 2024).

**Studying interpreted LSF** As the first interpreted LSF dataset of this scale, Matignon-LSF can be used to study the specificity of interpreted LSF in comparison with the original LSF that can be observed in other corpora. For example, the work of (Belissen et al., 2020) could be used to quantify the distribution of sign types in this dataset.

## 6. Conclusion

In this paper, we presented Matignon-LSF, a new dataset completely open to both research and private use. We gave an overview of the dataset and then presented the processing steps we applied for the collection and preparation.

The scripts we developed are publicly available so that they may be used to extend the dataset as new videos are produced and published every week. We also aim at adding other videos such as President or Prime Minister solo intervention. The corpus itself will be soon made available on the [Ortolang](#) platform.

This dataset is the first dataset of interpreted LSF, also usable outside public research. Future work should focus on aligning this dataset, in particular to facilitate the suggested perspectives.

## Acknowledgements

This work has been partially funded by the Bpifrance investment “Structuring Projects for Competitiveness” (PSPC), as part of the [Serveur Gestuel](#) project.

## Authors details

None of the authors are deaf. A deaf colleague, specialist in motion capture and virtual signer animation, belongs to our team but didn’t participate to this project. Moreover, we often collaborate with the Deaf community.



## Bibliographical References

- S. Albanie, G. Varol, L. Momeni, T. Afouras, Joon S. Chung, N. Fox, and A. Zisserman. 2020. [Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues](#). In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.
- S. Albanie, G. Varol, L. Momeni, H. Bull, T. Afouras, H. Chowdhury, N. Fox, B. Woll, R. Cooper, A. McParland, et al. 2021. [BOBSL: BBC-Oxford British Sign Language Dataset](#). In *ArXiv preprint*.
- A. Balvet, C. Courtin, D. Boutet, C. Cuxac, I. Fusellier-Souza, B. Garcia, M.-T. L’Huillier, and M. A. Sallandre. 2010. [The creagest project: a digitized and annotated corpus for french sign language \(Isf\) and natural gestural languages](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pages 469–475.
- V. Belissen, A. Braffort, and M. Gouiffès. 2020. [Experimenting the automatic recognition of non-conventionalized units in sign language](#). *Algorithms*, 13(12):310–336.
- A. Braffort. 2022. [Langue des signes française: Etat des lieux des ressources linguistiques et des traitements automatiques](#). In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 131–138. CNRS.
- H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman. 2021. [Aligning subtitles in sign language videos](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11552–11561.
- H. Bull, A. Braffort, and M. Gouiffès. 2020. [MEDI-API-SKEL -A 2D-Skeleton Video Database of French Sign Language With Aligned French Subtitles](#). In *12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6063–6068, Marseille, France.
- Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. 2018. [OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields](#). In *arXiv preprint arXiv:1812.08008*.
- J. Carreira and A. Zisserman. 2017. [Quo vadis, action recognition? a new model and the kinetics dataset](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308.
- D. Dayter. 2019. *Collocations in non-interpreted and simultaneously interpreted English: a corpus study*. Routledge.
- A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto. 2021. [How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language](#). In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2735–2744.
- H. Hu, W. Zhao, W. Zhou, and H. Li. 2021. [Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding](#). In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11087–11096.
- C. Huang, Joseph L. Mundy, and Charles A. Rothwell. 1994. [Model supported exploitation: Quick look, detection and counting, and change detection](#). In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 144–151.
- O. Koller, J. Forster, and H. Ney. 2015. [Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers](#). *Computer Vision and Image Understanding*, 141:108–125.
- M. Kopf, M. Schulder, and T. Hanke. 2023. [The sign language dataset compendium](#).
- J. Lascar, M. Gouiffès, A. Braffort, and C. Danet. 2024. [Annotation of Isf subtitled videos without a pre-existing dictionary](#). In *Workshop on the Representation and Processing of Sign Languages at the International Conference on Language Resources and Evaluation (sign-lang@LREC)*.
- C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. 2019. [Mediapipe: A framework for building perception pipelines](#). *CoRR*, abs/1906.08172.
- M. Müller, M. Alikhani, E. Avramidis, R. Bowden, A. Braffort, N. Cihan Camgöz, S. Ebling, C. España-Bonet, A. Göhring, R. Grundkiewicz, M. Inan, Z. Jiang, O. Koller, A. Moryossef, A. Rios, D. Shterionov, S. Sidler-Miserez, K. Tissi, and D. Van Landuyt. 2023. [Findings of the second WMT shared task on sign language translation \(WMT-SLT23\)](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.
- Y. Ouakrim, D. Beautemps, M. Gouiffès, T. Hueber, F. Berthommier, and A. Braffort. 2023. [A](#)

- multistream model for continuous recognition of lexical unit in french sign language. In *29<sup>o</sup> Colloque sur le traitement du signal et des images*", 2023-1182, pages 461–464. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.
- Y. Ouakrim, H. Bull, M. Gouiffès, D. Beautemps, T. Hueber, and A. Braffort. 2024. *Mediapi-RGB: Enabling technological breakthroughs in french sign language (LSF) research through an extensive Video-Text corpus*. *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2:139–148.
- K. Renz, N. C. Stache, S. Albanie, and G. Varol. 2021. *Sign language segmentation with temporal convolutional networks*. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2135–2139. IEEE.
- C. Stone and D. Russell. 2011. *Interpreting in international sign: decisions of deaf and non-deaf interpreters*. In *Proceedings of World Association of Sign Language Interpreters Conference*.
- L. Tarrés, G. I. Gállego, A. Duarte, J. Torres, and X. Giró-i Nieto. 2023. *Sign language translation from instructional videos*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5634.
- D. Uthus, G. Tanzer, and M. Georg. 2024. *Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus*. *Advances in Neural Information Processing Systems*, 36.
- G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman. 2021. *Read and attend: Temporal localisation in sign language videos*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16857–16866.
- L. Ventura, A. Duarte, and X. Giró-i Nieto. 2020. *Can everybody sign now? exploring sign language video generation from 2d poses*. *arXiv preprint arXiv:2012.10941*.
- Bull, H. and Braffort, A. and Gouiffès, M. 2019. *Mediapi-Skel corpus*. ISLRN 184-726-682-550-4.
- Bull, H. and Ouakrim, Y and Lascar, J. and Braffort, A. and Gouiffès, M. 2024. *Mediapi-RGB corpus*. ISLRN 421-833-561-507-6.

## Language Resource References

- Belissen, V. and Braffort, A. and Gouiffès, M. 2020. *Dicta-Sign-LSF-v2 corpus*. ISLRN 442-418-132-318-7.

# Phonological Transcription of the *Canadian Dictionary of ASL* as a Language Resource

Kathleen Currie Hall, Anushka Asthana, Maggie Reid,  
Yiran Gao, Grace Hobby, Oksana Tkachman, Kaili Vesik

University of British Columbia  
2613 West Mall, Vancouver, BC V6T 1Z4 Canada  
kathleen.hall@ubc.ca

## Abstract

This paper introduces the ongoing project of digitizing and phonologically transcribing the *The Canadian Dictionary of ASL* (Bailey and Dolby, 2002) to be used as a language resource. We describe the contents of the dictionary and the procedure used for creating the transcribed version, using the Sign Language Phonetic Annotator-Analyzer software (Hall et al., 2022). We also outline the benefits of creating a resource with such a detailed representation of the formational structure of signs.

**Keywords:** dictionary, phonological transcription, American Sign Language, Canada

## 1. Introduction

In this paper, we introduce an ongoing project to digitize and phonologically transcribe *The Canadian Dictionary of ASL*<sup>1</sup> (Bailey and Dolby 2002; henceforth *CD-ASL*), currently available in print only, as a language resource for phonological analysis. As Morgan (2022) says, “a digitized record of the formational content of signs that is easy to query on demand” is necessary for doing fine-grained, careful phonological analysis of sign languages (p. 99). Such a record facilitates, for example, the finding of minimal pairs, the understanding of the lexical frequency of different phonological parameters, the ability to analyse phonotactic restrictions, and more generally, the synthesis of phonetic and phonological information in a practical way. Digital records of the form of signs are also helpful for non-researchers, e.g., teachers or learners of a sign language who want to be able to look up a sign based on its formational characteristics rather than its gloss into a relevant spoken language. It is in this context that we have undertaken a detailed transcription of the *CD-ASL*.

### 1.1. Motivations

The widespread availability of digital tools allows for the creation of sign-language resources on a scale and with functionality that was impossible in previous years. However, much research effort has been invested in creating analog sign-language resources such as the dictionary we are using, and one of our aims is to help preserve

the valuable information in such documents for future use. Future use, however, requires that resources be readily available and easy to interact with. The *CD-ASL* is similar to most paper-based sign-language dictionaries in that it is organized by its English glosses rather than by any sign-language-specific feature such as phonological parameters. Thus, the user interested in signs that share a specific phonological trait (e.g., a specific handshape) is faced with a daunting task of manually sifting through the entire dictionary in search of such signs. Part of our motivation, then, is to create a freely available digital resource that will allow for phonologically based searching.

Most lexical databases of a sign language do provide some phonological information. As technology and research have progressed, however, more and more such information can be added, and we also see ourselves as contributing to the next stage of this endeavour. For example, the *ASL-LEX* database (Sehyr et al. 2021, Casselli et al. 2021), while extraordinarily useful in the breadth of information it covers, collapses certain phonological categories in ways that make answering some basic questions difficult. For instance, there is no way to easily search for a sign based on the number of syllables it contains. While signs are coded for repetition, which may be repetition of either a major or a minor movement in a sign, only the former would be thought to license multiple syllables. As another example, ‘contact’ in *ASL-LEX* is given only binary status, with no ability to search for what elements are in contact, when the contact happens, or what type of contact it is (e.g. continuous or holding, cf. Friedman 1976). *ISL-LEX* (Morgan et al., 2022b), on the other hand, which was built after the initial version of *ASL-LEX*, does include explicit information about syllables and con-

---

<sup>1</sup>ASL here is American Sign Language, the name of the sign language used in most parts of English-speaking Canada; see §2.

tact types. However, it still combines other categories, such as having a generic ‘combination’ category for orientation movement types instead of a compositional option to search by different specific combinations. We applaud all of these efforts to document phonological information and aim to build on the knowledge and experience of these projects, adding more detail as it becomes clear which information would be useful. The more languages that have documentation of phonological structure, the better our descriptions and theories of sign language phonology can be.

To these ends, we describe our ongoing project to provide a detailed phonological transcription of the signs from the *CD-ASL*, using software designed to facilitate such transcription of any sign language, Sign Language Phonetic Annotator-Analyzer software (Hall et al., 2022). The following sections describe the general contents of the dictionary (§2), the software and transcription procedures (§3), and the current state of the project and our initial examples of uses for the end product (§4). Before we do that, however, we believe it is important to be explicit about our own positionality with respect to this project.

## 1.2. Positionality

First, it is important to be transparent about the fact that all of the co-authors on this paper are hearing, and none of us is a fluent ASL signer. Most of us have taken a number of ASL courses, all of which have been taught by Deaf signers who also emphasize awareness of Deaf culture and communities.

We recognize that the lack of Deaf signers as primary researchers on the project is a significant shortcoming for both practical and social reasons. At the same time, we think that it is important for researchers at spoken-language-biased institutions, such as the University of British Columbia, where we are based, not to ignore sign languages simply because our systems are not yet designed to fully support d/Deaf students and colleagues (and we are independently involved in trying to change that). We have made efforts to collaborate at every stage of this project with Deaf signers to ensure that the project is one that is generally supported by the Deaf community and that we are transcribing signs accurately.

This overall situation is indeed one of the reasons we chose to transcribe the *CD-ASL* as a resource: it is seen as a valuable tool for Canadian signers, and much of the work that needs to be done to make it phonologically searchable is the ‘grunt work’ of simply taking the pre-existing textual descriptions and translating them into phonological transcriptions, a task that can be done by anyone who is trained, and for which we do not

have to overly burden community members with laborious tasks.

At the same time, there are many cases in which the dictionary text is underspecified and/or mismatches the image provided (e.g. as in Figure 1 for ADDRESS, discussed in §2). In these cases, we consult with a Deaf signer to clarify the correct baseline transcription to be used.

Here, we would like to directly acknowledge in the text of this paper the contributions of Deaf scholars and community members who have been directly consulted on this project, listed here in alphabetical order: Vincent Chauvet, Joanne Cripps, Leanne Gallant, Julie Hochgesang, Nigel Howard, Jonathan MacDonald, Gary Malkowski, and Erin Wilkinson. We owe them a debt of gratitude for helping us in our endeavours. Having said this, we also take full responsibility for any errors in our representations.

## 2. The Canadian Dictionary of ASL

The *CD-ASL* (Bailey and Dolby, 2002) was published in 2002 by the Canadian Cultural Society of the Deaf and University of Alberta Press to document the signs of American Sign Language (ASL) as they are used in Canada. Work started on the dictionary in 1982, and the form of signs in the dictionary therefore reflects ASL as it would have been most commonly used in the last two decades of the 20th century. As explained in the preface, “the *Dictionary* pays special attention to subjects of particular interest to Deaf Canadians—bilingual and bicultural education, residential schools, ice hockey and other winter sports, parliamentary government, weather and geographic features, historical events and geographic place names” (p. XI). The *CD-ASL* also has a special focus on the regional variation of signs across Canada, with variants from the Pacific (British Columbia), Prairie (Alberta, Saskatchewan, and Manitoba), Central (Ontario and Québec), and Atlantic (New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador) regions each being tagged in individual regional-specific entries.<sup>2</sup> Hence, this dictionary is unique in its documentation of *Canadian ASL* and allows research to be done looking at lexical form variation (cf. Stamp et al. 2014; Bayley et al. 2015; Palfreyman 2015; and Siu 2016, among others, for studies on variation in sign languages).

The *CD-ASL* contains over 8700 entries (see e.g., Figures 1 and 3), each typically given a definition in English, an English sample sentence, an English prose explanation of the formational structure of the sign, and a line drawing depicting the

---

<sup>2</sup>The three northern territories of Canada are noticeably absent from this tagging.



ASL sign. Some of these entries, however, are homophones rather than unique forms (e.g. ACCESS and ADMISSION are separate entries in the dictionary, but each is accompanied with the same description and image). Additionally, some of the entries are represented simply as fingerspelled words with no separate ASL form (e.g. AGENDA).

Within the description of the form, each handshape is given an absolute categorical label, aligned with the set of 114 handshapes identified by the editors of the dictionary as occurring in Canadian ASL. All other phonological information is described in prose and varies in terms of the consistency of information given with respect to palm orientation, location, movement, and non-manual characteristics. Occasionally, there is a mismatch between the prose description and the line drawing provided. An example entry with such a mismatch is shown in Figure 1, for the sign ADDRESS. Note that the text suggests a repeated straight upward movement, while the arrows in the image suggest that the movement is instead a circular action. While our internal convention is to prioritize the text over the image in such cases with our initial coding, we are also subsequently checking all such cases with a Deaf signer to resolve the conflict.

The 840-page *CD-ASL* is currently published only in a hardcover format (<https://ualbertapress.ca/9780888643001/the-canadian-dictionary-of-asl/>). As with all such paper-based resources, then, searching is difficult and entirely dependent on the organization of the written text. In this case, the entries are organized alphabetically by English gloss, such that searching by any phonological parameter (handshape, location, etc.) is entirely impossible. One of our goals in this work was to create a digitally accessible, phonologically organized resource that can be searched in this way. Details of our procedure for creating this resource are described in the next section.



**address:** *n.* postal designation or place of residence. *She put her new address on the envelope.*

SIGN #1: Horizontal 'EXTENDED A' hands, palms toward the body, are simultaneously brushed upward twice against the chest.

Figure 1: An example of an entry in the *CD-ASL*, for the sign ADDRESS.

### 3. Transcription Procedure

To create the digital version of the form-based entries, we are using the [Sign Language Pho-](#)

[netic Annotator/Analyzer](#) software (SLP-AA; Hall et al. 2022). This software is a free and open-source tool (<https://github.com/PhonologicalCorpusTools/SLPAA/>) designed to facilitate detailed form-based transcription of signs. Transcriptions are done through menus of pre-defined options. Approaching transcription this way has several advantages. First, text-based descriptions are more human-readable than many notation systems (see discussion in Hochgesang 2014), allowing transcribers to be trained more quickly and allowing non-trained users of the resource to more readily understand the transcriptions. Second, providing the options as pre-existing menu items preserves the utility of standardization of transcription and ease of computer-based searches for particular characteristics. An example of some of the options for coding path movements in SLP-AA is shown in Figure 2. Note that there are still places for users to enter their own text if needed—for example, if the shape of the movement is something other than one of the pre-specified ones. Currently, the software only presents these menu choices in English; this is a potential drawback for more widespread usage.

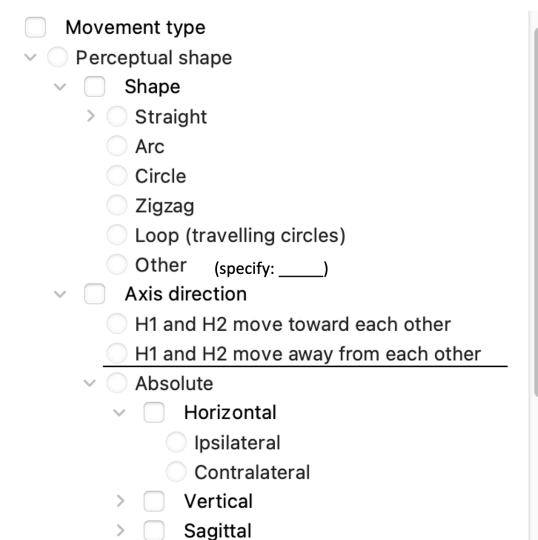


Figure 2: A screenshot of part of the movement selection options in the SLP-AA software.

This software is still under simultaneous development with the transcription of the *CD-ASL*, by an overlapping but not identical set of researchers, and the two endeavours are mutually beneficial. Using the software to transcribe actual forms allows us to improve the coverage and user interface of the software, and the existence of the software allows us to create standardized, searchable transcriptions of the entries in the *CD-ASL*.

### 3.1. Selection of Entries

Due to the simultaneous development of the SLP-AA software, we are approaching the transcription of the *CD-ASL* in stages. As a first pass, we are coding a representative sample of signs from the dictionary rather than immediately working on coding all of the entries. To provide us a concrete guideline for selection, we chose to select all entries from the *CD-ASL* that share a gloss with the entries in the ASL-LEX database (Casselli et al., 2021). This also allows for direct comparisons both between the actual signs (e.g., American vs. Canadian dialect differences) and between the phonological transcriptions of signs that happen to have similar forms. Note that we just use the glosses in ASL-LEX to select glosses from the *CD-ASL*; we do not filter signs by whether the actual forms are similar across the two sources. For example, if there are two separate entries in the *CD-ASL* for related but not-identical concepts (e.g., ADULT vs. ADULTS), we select for inclusion only the one for which there is an exact *gloss* match in ASL-LEX (in this case, ADULT). This is despite the fact that the form for ADULT in ASL-LEX happens to be more similar to the form for ADULTS in the *CD-ASL*.

Once a gloss has been selected, all of the various entries for that gloss from the *CD-ASL* are transcribed, such that in many cases, a single gloss from ASL-LEX results in multiple entries in our resource (e.g., PASS has five unique forms in the *CD-ASL*, representing six different semantic senses of the English word ‘pass’). At the same time, not every gloss that occurs in ASL-LEX occurs in the *CD-ASL*; such glosses are skipped (e.g., ACCENT occurs in ASL-LEX but there is no entry with this gloss in the *CD-ASL*). Occasionally, a gloss from ASL-LEX occurs under a different gloss in the *CD-ASL*, and such entries are also transcribed (e.g., the ASL-LEX gloss ACCOUNTANT is listed as the ‘same sign’ under the *CD-ASL* entry ACCOUNTING, and so ACCOUNTANT is included in our transcriptions).

### 3.2. Parameters and Other Phonological Content

When we began transcribing entries from the *CD-ASL* in January of 2023, the SLP-AA software supported coding the *sign type* of signs along with *handshape*, *movement*, and *location* specifications. All of our signs are coded for these parameters. In the fall of 2023, with developments in the SLP-AA software, we were able to start adding in what we refer to as *relation* elements, such as contact specifications and relative orientation; about half of our signs currently are coded for relation. Absolute orientation and non-manual parameters

are still being implemented in the software and have not been coded for any signs. Further explanation of how these parameters are coded follows immediately below; more complete descriptions are provided in Hall et al. (2022), and full documentation of the software and its choices for transcription is also under development.

#### 3.2.1. Sign Type

The sign type choices in SLP-AA roughly follow those laid out by Battison (1978). Rather than assigning explicit numbers to each type, however, the elements that determine a sign’s type are coded separately, again to allow for easier searching of specific characteristics. For example, the options in the sign type module allow a user to specify that a sign is one- or two-handed, and if it is two-handed, whether both hands move or only one, and if both hands move, whether they move similarly, etc. Transcribers base their selections on the text of the dictionary entry.

#### 3.2.2. Timing

One of the ways in which the SLP-AA transcriptions are more detailed than most other such notations is that they support full detail for indicating the relative timing of each parameter, even in a static resource such as a dictionary (as compared to a real-time resource like a corpus). For example, as mentioned above, ASL-LEX codes whether or not there is contact in a sign, but does not indicate when such a contact occurs during the sign or which elements make contact. In ISL-LEX (Morgan et al. 2022a, Morgan et al. 2022b), signs are explicitly allowed to have two path movements or two locations, each individually specified. To make timing even more flexible, in SLP-AA, each sign is assigned an abstract ‘x-slot’ structure, such that specific elements like contact, location, or movement, can be associated with points or intervals at any relevant time during the sign. For the *CD-ASL* coding, we define x-slots essentially as syllables, with each iteration of the largest movement within a sign defining a syllable and hence an x-slot (see e.g. Stack 1988; Wilbur 2011). A simple monosyllabic sign, then, will have a handshape and location defined at the beginning of an x-slot, then have a movement that lasts the entire x-slot, and a new handshape and/or location defined at the end of the x-slot, depending on what has changed. If the movement changes only the handshape, the location is assigned to have the same duration as the whole x-slot, and vice versa. For example, Figure 3 shows the dictionary entry for the monosyllabic sign RED, and Figure 4 shows the resulting summary of the transcription in SLP-AA. The sign type is shown across the top, spanning one x-slot,

and modules for movement, location, relation, and hand configuration are assigned to their relative timing. In this case, the movement and location each last for the entire x-slot duration, the hand configuration is different at the beginning and end, and a relation module is used only at the beginning.



**red:** *adj.* the colour of blood. *He wore a red shirt and white shorts for Canada Day.*

**SIGN:** Vertical right **'ONE'** hand is held with palm facing the body and tip of forefinger touching the lower lip. As the hand is then drawn very firmly forward at a downward angle, the forefinger crooks to form an **'X'** shape.

Figure 3: An example of an entry in the *CD-ASL*, for the sign RED.

### 3.2.3. Handshape

As mentioned in §2, the *CD-ASL* provides a categorical label for each handshape used in the dictionary, along with images of each canonical version of the handshape and descriptions of their conventionalized labels such as 'clawed' or 'spread.' Each of the handshapes that is included in the *CD-ASL* has been pre-transcribed as a 'pre-defined' handshape within the SLP-AA software, using the Johnson and Liddell (2011a,b, 2012) transcription system, modified as described in Tkachman et al. (2016). Thus, for each sign being transcribed, the transcriber only has to select the relevant handshape name and associate it to the appropriate timepoints in the sign. For example, for the sign RED, shown in Figure 3, the transcriber would select "ONE" and associate this with the beginning of the x-slot. This associates both the phonological handshape label and the detailed phonetic transcription of this hand configuration with this sign; both are shown in the tooltip obtained by hovering over the first hand configuration element, as shown in Figure 4. A similar process is used to transcribe the "X" handshape at the end of the sign.

### 3.2.4. Movement

Movements in the text of the *CD-ASL* are described in prose. While there are some terms that are used repeatedly (such as "move alternately," or "brought together," or "circle"), there is much variability in the specific wording. One of the advantages of using the SLP-AA software to transcribe the dictionary is to standardize these descriptions, such that users can easily search for

or calculate the frequency of particular types of movement. Transcribers 'translate' the prose descriptions into the pre-set parameter values within the software. These parameter values are largely derived from classic phonological descriptions of movement, focusing on shapes / path movements, joint-specific internal movements, and what is often referred to as 'manner' of movement, e.g. directionality, repetition, and other specific characteristics like increased force or speed (e.g. discussion in Brentari 1998; van der Kooij 2002; Sandler and Lillo-Martin 2006; Sandler 2011; Morgan 2022).

For example, in RED, there are two simultaneous movements, one that would typically be described as a 'path' movement, where the hand moves "very firmly" in a straight line forward and away from the signer, and one that involves the index finger "crook"-ing (called 'hooking' in SLP-AA). Each of these movements is fully transcribed with a separate instance of a movement module in SLP-AA, and associated with the entire x-slot (these are shown as H1.Mov1 and H1.Mov2 in Figure 4). One convention we use here is that if the text entry does not specify whether the movement is a path movement or a joint-specific / local movement, we default to the path interpretation, and this is another type of information that we consult with a Deaf signer about.

Sometimes, instead of using explanatory notes, the dictionary provides a special symbol to mark a key aspect of a sign's production. One example is directional verbs, i.e., verbs that may move in different positions in signing space, depending on where the positions of people in the communicative context are. Such signs are marked with a special symbol that indicates their nature as directional. Our internal convention is that our basic transcription follows the baseline information in the text about the direction of the sign's movement, but we also mark such signs as directional verbs in the coding, such that they could all be found in a subsequent search if desired.

### 3.2.5. Location

As with movements, locations are described in prose in the dictionary and are translated into standardized SLP-AA terminology. In the software, there are two basic choices for location types: signing space locations, designated by locations on the horizontal, vertical, and sagittal axes, and body locations. The choices for body locations are essentially a super-set of the locations in Brentari (1998); Hanke (2004); Johnson and Liddell (2021) and Morgan (2022).

In RED (Figure 3), for example, "the lower lip" is translated into the SLP-AA specification of being a body location of the 'lower lip,' which is hi-

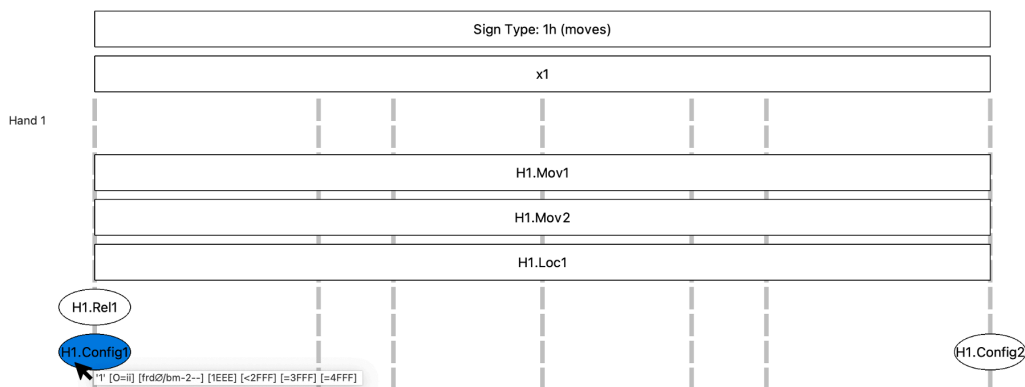


Figure 4: The SLP-AA summary window for the sign RED. Each element in the summary can be clicked to show the complete coding; hovering over an element gives a preview. Here, the first hand configuration (for the “ONE” handshape) is selected, and a preview of the full phonetic transcription is shown.

erarchically nested under ‘Head / Face / Mouth / Lips.’ A user could use any of these higher-level categories instead; to code the *CD-ASL*, we use the categories that most closely align to the text description. The details of contact are specified as part of the relation module, as described in the next section.

As with other parameters, we have certain conventions that allow us to code otherwise underspecified signs. For example, most one-handed signs, especially those in neutral space, are not actively specified in the text as occurring on one side of the body or the other. We default to assuming that one-handed signs are on the ipsilateral side of the body, but if there is any reason to suspect that a particular sign is not so located (e.g., the accompanying image shows the hand in a different location), we would ask a Deaf signer consultant about the typical production.

### 3.2.6. Relation

The final type of information currently being included in the transcribed *CD-ASL* is what we call ‘relation’ information.<sup>3</sup> This includes all types of relations between two elements, such as the relation between the two hands or between one or both hands and a particular location or movement. This can be used to code spatial relations (e.g., Hand 1 is above and in front of Hand 2), presence or absence of contact (e.g., Hand 1 contacts Hand 2), type of contact (e.g., the contact between Hand 1 and Hand 2 is ‘holding’ or ‘continuous,’ cf. [Friedman 1976](#)), distance (e.g., the hands are close to or far from a location), and the hand part that is

<sup>3</sup>Absolute orientation, which we take to be all statements of “palm facing” directions in the dictionary, e.g. “palm facing the body” in the entry for RED in Figure 3, can also be coded with SLP-AA, but we have not yet invested resources into doing this coding, instead prioritizing relative orientation.

relevant to a movement or location (e.g., the ulnar side of Hand 1 leads a movement or makes contact with a location; cf. relative orientation as discussed in [Crasborn and van der Kooij 1997](#)).

In RED (Figure 3), the fact that it is the “tip of the forefinger” that touches the lower lip at the beginning of the sign is coded as a relation module that is specifically linked to the location module. This relation module marks that Hand 1—and specifically, the tip of the index finger—has contact with this lower lip location at the beginning of the x-slot. As with other parameters, any ambiguities or underspecifications are checked with a Deaf signer.

### 3.3. Updating Dictionary Entries

As noted above, we are in the process of verifying underspecified and conflicting entries with a Deaf signer to make sure our entries are as accurate as possible. Our consultant points out multiple kinds of issues with the current dictionary entries, including both entirely out-of-use signs and individual elements of the production of signs that do not match current usage. We are currently only modifying the *CD-ASL* entries where they were underspecified or self-conflicting, rather than actively changing entries to be more modern. Digitizing older sign language dictionaries at the level of phonetic and phonological detail like ours enables researchers to ask meaningful questions about language change and language evolution, e.g., how more gestural elements of sign-language communication become grammaticalized, reduced, etc. (cf. [Shaffer and Janzen 2000](#); [Janzen and Shaffer 2002](#)). At the same time, we are keeping track of all such additional information provided by our consultant, so that we can cross-check with other Deaf signers and potentially provide modern equivalents to dictionary entries in the future.



## 4. Findings and Future Studies

As of the time of submission, approximately 2000 signs from the *CD-ASL* have been transcribed, with transcribers currently working on the letters *P* and *R*. These are all unique forms; signs that have separate entries but are repeated forms from earlier entries have not yet been included, as these will eventually be single entries tagged with multiple glosses. However, the ~2000 signs do include multiple different forms for the same gloss (e.g., including both the generic and the Atlantic Canadian forms of the sign ADDRESS ‘postal designation’ as well as the different ASL forms used for ADDRESS ‘postal designation’ vs. ADDRESS ‘lecture’). Transcribed signs also exclude labelled compound signs (e.g., ABNORMAL, described as “ASL concept NOT - NORMAL”) but include finger-spelled signs (approx. 300 signs).

When complete, the transcribed version of the dictionary will be made publicly available as a binary .slpaa file, which is the specific file type that can be read and interpreted by the SLP-AA software. We are also actively developing the “Analysis” component of the software to allow for ease of searching and comparison of signs. We are hoping to also distribute a less software-dependent version of the transcribed signs, e.g. as a .csv, a .json, or a .sql file, depending on the complexity of the data structures involved.

This work in progress has allowed us to have useful insights into phonological description and structure, even before we have a fully complete dictionary resource. For example, we have been forced to confront the difficulty of handling circular direction terms in a way that is consistent and searchable. The *CD-ASL* assumes a right-handed signer, but we would like our resource to be usable by and relatable to all signers, regardless of hand dominance. Furthermore, the dictionary is inconsistent in how it describes circular directions even for a right-handed signer, sometimes adopting the perspective of the signer and sometimes the addressee, and sometimes not specifying the perspective. To create a consistent, inclusive, and searchable record of these signs, we have adapted the coding conventions away from terms like “clockwise” and “counter-clockwise” and instead use terms like “ipsilateral from the top of the circle” (where the “top” is conventionally defined to be the highest point for circles on the vertical and sagittal planes and the most distal point for circles on the horizontal plane). We hope that an update like this might be extended to other descriptive projects to facilitate cross-resource comparison as well.

Another future direction that this project has already suggested is the investigation of the fore-

arm in lexical specification. There have been a number of signs in the *CD-ASL* whose descriptions make it clear that the position of the forearm was deemed important to the writers of the dictionary. The potential relevance of the forearm has been noted since at least Stokoe et al. (1965), where certain signs were said to involve a “prominent” use of the forearm of the dominant hand, e.g. in the sign DAY (<https://www.handspeak.com/word/537/>; Lapiak 1995). Stokoe’s notation convention was to include a checkmark for such signs, and Johnson and Liddell (2012) adopt the same convention in their phonetic notation system. However, there are a wide variety of actual cases in which forearms may be relevant. Compare, for example, DAY to the sign for CASTLE as described in the dictionary, which is similar to the version marked ‘regional variation’ at <https://www.handspeak.com/word/1723/> (Lapiak, 1995). This sign involves both forearms resting horizontally one on top of the other at the beginning of the sign and each being raised vertically at the end of the sign. Another potential use for the forearm is as in BARK (as in ‘tree bark’) and BRIDGE, where the forearm of the non-dominant hand is used as an iconic location for the dominant hand to act upon. Only by having a detailed phonological transcription of signs in a language—specifically, detailed enough to include information about forearm position and movement—can we hope to catalogue, classify, and eventually fully understand the phonological role of the forearm as an articulator in sign languages.

There are many such specific examples that arise as we code, even when we limit ourselves to the glosses that also occur in ASL-LEX. While we recognize that many early efforts to create databases for sign languages have focused for good reason on the most canonical types of signs, we think that the field is in a position to dive more deeply into these less prototypical types of signs and include them in our phonological research.

## 5. Conclusion

We see digitizing older sign-language resources such as the *CD-ASL* as a way to acknowledge past signers and past research, and as a means of beginning to address more detailed and specific questions of diachronic change and synchronic phonological structure. We believe that transcribing signs on a more detailed level than has previously been possible will provide us with much greater insight into the phonological systems in sign languages. Having a digitized and freely available resource of this nature should also be helpful to Canadian ASL users who are trying to interact directly with the formational structure of the

language and not through its English translations. We hope that our experience with digitizing the CD-ASL will also inspire other researchers to digitize dictionaries of other sign languages, regardless of their publication date, and to create both lexical and corpus resources that include a fine-grained level of phonological detail.

## 6. Acknowledgements

This research is supported in part by funding from the Social Sciences and Humanities Research Council. We are also grateful to numerous contributors and advisors on this project, including: Leanne Gallant and the Canadian Cultural Society of the Deaf, Douglas Hildebrand and University of Alberta Press, Yurika Aonuki, Ashley Chand, Vincent Chauvet, Sophie Cook, Joanne Cripps, Brian Diep, Michael Fry, Paris Gappmayr, Julie Hochgesang, Nigel Howard, Shannon Hsu, Janet Jamieson, Cristina Lee, Roger Yu-Hsiang Lo, J. Scott Mackie, Jonathan MacDonald, Gary Malkowski, Natalie Michaelian, Hope Morgan, Stanley Nam, April Poy, Nathan Sanders, Nico Tolmie, Erin Wilkinson, and Grace Zhang. All errors are our own.

## 7. Bibliographical References

- Carole Sue Bailey and Kathy Dolby, editors. 2002. *The Canadian dictionary of ASL*. The University of Alberta Press.
- Robbin Battison. 1978. *Lexical borrowing in American Sign Language*. Linstok Press, Silver Spring, MD.
- Robert Bayley, Adam Schembri, and Ceil Lucas. 2015. *Variation and change in sign languages*, chapter 4. Cambridge University Press.
- Diane Brentari. 1998. *A prosodic model of sign language phonology*. MIT Press, Cambridge, MA.
- Onno Crasborn and Els van der Kooij. 1997. *Relative orientation in sign language phonology*. *Linguistics in the Netherlands*, pages 37–48.
- Lynn Alice Friedman. 1976. *Phonology of a soundless language: phonological structure of the American Sign Language*. Ph.D. thesis, University of California, Berkeley.
- Kathleen Currie Hall, Yurika Aonuki, Kaili Vesik, April Poy, and Nico Tolmie. 2022. *Sign language phonetic annotator-analyzer: Open-source software for form-based analysis of sign languages*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 59–66, Marseille, France. European Language Resources Association (ELRA).
- Thomas Hanke. 2004. *HamNoSys – representing sign language data in language resources and language processing contexts*. In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).
- Julie A. Hochgesang. 2014. Using design principles to consider representation of the hand in some notation systems. *Sign Language Studies*, 14(4):488–542.
- Terry Janzen and Barbara Shaffer. 2002. *Gesture as the substrate in the process of ASL grammaticalization*, pages 199–223. Cambridge University Press.
- Robert E. Johnson and Scott K. Liddell. 2011a. *A segmental framework for representing signs phonetically*. *Sign Language Studies*, 11(3):408–463.
- Robert E. Johnson and Scott K. Liddell. 2011b. *Toward a phonetic representation of hand configuration: the fingers*. *Sign Language Studies*, 12(1):5–45.
- Robert E. Johnson and Scott K. Liddell. 2012. *Toward a phonetic representation of hand configuration: the thumb*. *Sign Language Studies*, 12(2):316–333.
- Robert E. Johnson and Scott K. Liddell. 2021. *Toward a phonetic description of hand placement on bearings*. *Sign Language Studies*, 22(1):131–180.
- Els van der Kooij. 2002. *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. Ph.D. thesis, Leiden University.
- Hope E. Morgan. 2022. *A phonological grammar of Kenyan Sign Language*. De Gruyter Mouton.
- Hope E. Morgan, Wendy Sandler, Rose Stamp, and Rama Novogrodsky. 2022a. *ISL-LEX v.1: An online lexical resource of Israeli Sign Language*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 148–153, Marseille,

France. European Language Resources Association (ELRA).

Nick Palfreyman. 2015. *Sign language varieties of Indonesia: a linguistic and sociolinguistic investigation*. Ph.D. thesis, University of Central Lancashire.

Wendy Sandler. 2011. *The phonology of movement in sign language*, chapter 24. Wiley-Blackwell, Oxford.

Wendy Sandler and Diane Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge UP, Cambridge.

Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. [The ASL-LEX 2.0 project](#). *The Journal of Deaf Studies and Deaf Education*, 26(2):263–277.

Barbara Shaffer and Terry Janzen. 2000. Gesture, lexical words, and grammar: Grammaticalization processes in ASL. *Annual Meeting of the Berkeley Linguistics Society*, 26(1):235–245.

Sign Language Phonetic Annotator/Analyzer. [[computer software](#)].

Way Yan Rebecca Siu. 2016. *Sociolinguistic variation in Hong Kong Sign Language*. Ph.D. thesis, Te Herenga Waka–Victoria University of Wellington.

Kelly Magee Stack. 1988. Tiers and syllable structure in American Sign Language: Evidence from phonotactics.

Rose Stamp, Adam Schembri, Jordan Fenlon, Ramas Rentelis, Bencie Woll, and Kearsy Cormier. 2014. [Lexical variation and change in British Sign Language](#). *PLOS One*, 9(4):e94053.

William C. Stokoe, Dorothy C. Casterline, and Carl G Croneberg. 1965. *A dictionary of ASL on linguistic principles*. Linstok Press, Silver Spring, MD.

Oksana Tkachman, Kathleen Currie Hall, André Xavier, and Bryan Gick. 2016. [Sign language phonetic annotation meets phonological corpus-tools: Towards a sign language toolset for phonetic notation and phonological analysis](#). In *Proceedings of the 2015 Annual Meeting on Phonology*.

Ronnie Wilbur. 2011. *Sign syllables*, chapter 56. Wiley Blackwell, Oxford.

## 8. Language Resource References

Casselli, Naomi and Emmorey, Karen and Sehyr, Zed Sevcikova and Cohen-Goldberg, Ariel and O'Grady Farnady, Cindy. 2021. *ASL-LEX 2.0: Visualizing the ASL lexicon*. PID <https://asl-lex.org/>.

Lapiak, Jolanta. 1995. *HandSpeak®*. PID <https://www.handspeak.com/>.

Morgan, Hope and Sandler, Wendy and Novogrodsky, Rama. 2022b. *ISL-LEX 1.0: A Database of Phonological and Lexical Properties for 961 Signs in Israeli Sign Language*. PID <https://sites.google.com/view/isl-lex>.

# Retrospective of Kazakh-Russian Sign Language Corpus Formation

Alfarabi Imashev<sup>1</sup> , Aigerim Kydyrbekova<sup>1</sup>, Medet Mukushev<sup>1</sup>,  
Anara Sandygulova<sup>1</sup>, Shynggys Islam<sup>2</sup>, Khassan Israilov<sup>3</sup>,  
Aibek Makazhanov<sup>2</sup>, Zhandos Yessenbayev<sup>2</sup>

<sup>1</sup>Department of Robotics Engineering, Nazarbayev University, Kazakhstan

<sup>2</sup>Computer Science Lab, National Laboratory Astana, Nazarbayev University, Kazakhstan

<sup>3</sup>Public Association “Kazakh Deaf Society”, Astana branch, Kazakhstan

{alfarabi.imashev, aigerim.kydyrbekova, mmukushev, anara.sandygulova}@nu.edu.kz,  
sislam@alumni.nu.edu.kz, {aibek.makazhanov, zhyessenbayev}@nu.edu.kz

## Abstract

Sign language (SL) is a mode of communication that, in most cases, relies on visual perception exclusively and uses the visual-gestural modality. The advent of machine learning techniques has expanded the range of potential applications, not only in industry but also in addressing societal needs. Previous research has already demonstrated encouraging outcomes in developing sign language recognition systems that are both quite accurate and resilient. Nevertheless, the effectiveness and use of algorithms are impacted not only by their accessibility but also, at times to a greater extent, by the presence of substantial quantities of pertinent data. At the start of the local sign language corpus collection in 2015, there was a notable deficit of local Kazakh-Russian Sign Language data available for computer vision and machine-learning tasks. There were already corpora of another lexically close language, Russian Sign Language, but they were aimed at and tailored for linguistic research. We initiated the procedure by collecting data appropriate for machine-learning purposes. The subsets have been incorporated into the principal corpus and will be subject to future enhancements and refinements. This paper provides an overview of the collected components of the Kazakh-Russian Sign Language Corpus and the resulting outcomes derived from them.

**Keywords:** sign language, dataset collection, overview

## 1. Introduction

The emergence of machine learning approaches and techniques has broadened the scope of possible applications, not just in business or industry but also in meeting social demands. Previous research undertaken before 2015 has already shown promising results in the development of sign language recognition systems that are both highly accurate and durable. However, the efficiency and use of algorithms are influenced not only by their availability but also, often to a greater extent, by the existence of significant amounts of relevant data.

The government of Kazakhstan offers each deaf individual 60 hours per year of free sign language interpreting service support. These hours can be spent on medical, legal, or other communication requirements. The scarcity of interpreters per capita and the lack of remunerated interpreting services raise the necessity of supplementary alternative instruments for sign language recognition, translation and generation, which require datasets to train on. Regrettably, in 2015 there was not any dataset on local Kazakh-Russian Sign Language (K-RSL); there were corpora of similar Russian Sign Language (RSL) from Novosibirsk and Saint-Petersburg, but they were focused on linguistic research.

Thus, we decided to start collecting relevant data of local K-RSL suitable for machine learning applications. The sign language used by the deaf signers' community in Kazakhstan is not indigenous

and is closely related to RSL as well as other sign language within the CIS. All of these sign languages have their roots in the Soviet Union's centralized language policy, which established the signing system. While no formal study comparing K-RSL with RSL was conducted, the expertise of interpreters, and our observations indicate a significant similarity in vocabulary and frequent mutual intelligibility.

Nevertheless, the deaf community in Kazakhstan has already assimilated distinctive and unique themes into the local sign language, such as native musical instruments, regional delicacies, famous sites, significant figures, traditional beliefs, and more. Note that although RSL and K-RSL share many lexical similarities, it is uncertain if this extends to other linguistic aspects of both languages.

This paper provides a concise overview of the collected components of the Kazakh-Russian Sign Language Corpus aiming at applying machine learning approaches, and the resulting outcomes derived from them within the last decade.

The following section provides brief overview on related datasets existed in 2015. Section 3 offers a summary of subsets present in the current corpus, focused on several linguistic properties often seen in most sign languages, such as phonological minimum pairings, sign variability, and sign polysemy. Section 4 explores potential alternative methods for acquiring new types of sign language datasets.



Table 1: Hand Image Datasets (VS: Vocabulary Size, NP: Number of Participants)

Dataset	Volume	VS	NP	Resolution
NUS-I (Kumar et al., 2010)	480	10	24	160x120
NUS-II (Pisharady et al., 2013)	2000+750	10	40	160x120, 320x240
Polish Sign Language - I (Kawulok et al., 2013)	899	25	12	174x131 to 640x480
Polish Sign Language - II	85	13	3	4672x3104
Polish Sign Language - III	574	32	18	3264x4928
ASL Finger Spelling Dataset (Pugeault and Bowden, 2011)	48,000	24	5+4	128x128
J. Triesch I (Triesch and Von Der Malsburg, 1996)	720	10	24	128x128 (gray, 8 bit)
J. Triesch II (Triesch and Von Der Malsburg, 2001)	1000	12	19	128x128 (color)
MU ASL dataset (Barczak et al., 2011)	2524	36	5	high-res

Table 2: Video Datasets (VS: Vocabulary Size, NP: Number of Participants)

Dataset	Volume	VS	NP	Resolution
ASLLVD (Neidle et al., 2012)	9,800 tokens	3,300	1-6	640x480, 60fps 1600x1200, 30fps
BosphorusSign22k (Özdemir et al., 2020)	22,542 (19h)	744	6	1920x1080, 30fps
CSL-1 (Huang et al., 2018)	25,000 (100h)	178	50	1920x1080, 25 fps
RWTH-PHOENIX-Weather (Forster et al., 2012)	21,822 (1,980 sent.)	911	7	210x260, 25 fps
Purdue RVL-SLLL (Martinez et al., 2002)	2,576	39	14	640x480
DEVISIGN (Chai et al., 2014)	24,000(21.87h)	2000	30	640x480
SIGNUM (Von Agris et al., 2008)	33,210 (55.3h)	450, 780	25	776x578, 30 fps
RWTH-BOSTON (Athitsos et al., 2008)	843	406	5	324x242, 30 fps
DGS - KORPUS (Nishio et al., 2010)	50h (public)	530	330	640x360, 50 fps

## 2. Related Work

The task of finding a database that is optimal for machine learning and creating a model is specific and individual, for each particular task posed by the researcher. At the beginning of the study, we encountered several dataset containing images of the hands. We mostly did not take into account datasets designed for Kinect or Key-glove like devices, as they do not fulfill the necessary criteria of our goal, which is the ability of the system to operate with K-RSL without the need for any extra costly technological equipment. After reviewing which ML algorithms to test, we decided to revise the following image (see Table 1) and video (see Table 2) datasets available to figure out the best practices of dataset collection taking place at that moment (before 2015 and in 2020).

## 3. Collected Datasets

This section provides a brief account of the progressive growth of the K-RSL corpus, encompassing all datasets gathered for it from 2015 until the present day.

At the outset of our research, none of the sign language datasets mentioned in the literature followed any strict established requirements for recognizing continuous sign language that is not dependent on a signer. In contrast to voice recognition, there was no pre-existing standard, baseline, or reference point. Therefore, we have tried to collect a dataset that is anticipated to assist re-

search efforts for scholars who exhibit interest in the sign language recognition area. We believe that this dataset has the potential to become a benchmark for researchers who are studying advanced sign language recognition algorithms. It is signer-independent and suitable for continuous recognition. Furthermore, it includes cases of sign variability, polysemy (where the meaning of a sign is determined by mouthings), and phonological minimal pairs, which are very similar in performance. These factors make the task of automatic recognition more challenging and increase the complexity of the problem.

It is noteworthy that the deaf and hard-of-hearing community in Kazakhstan exhibits a high degree of insularity. Regrettably, according to Kazakhstan Deaf Society authorities and interpreters' experience, these issues arose due to instances of fraudulent activities perpetrated against individuals, including internet fraud, property crimes, violations of contracts, and lower wages, along with several instances of being involved in sects. All these negative experiences were deposited in memory and deeply ingrained in the local deaf culture, as was evident in how they viewed all outsiders. This led to the situation where interpreters and the state or non-profit deaf organizations became the primary conduits for establishing first communications and collaborations.

At the moment when our research began, there was a dominance of descriptors and feature extraction approaches in computer vision, and therefore, we also relied on the well-known ones and could

cooperate with four sign language interpreters only for our first attempt.

One major limitation of the sign language recognition field, when we started our research, was that all trustworthy and reputable video data sources consisted of video data, which was entirely created in a controlled “laboratory” setting. In such settings, the camera remains stationary, the background is uniform and consistent, and the lighting conditions are usually predetermined and unchanging. This was the reason why we decided to collect 1/3 of our first dataset outside the lab (Figure 2).

Based on previous linguistic and applied research, as well as the increasing availability of technologies that can extract coordinates of the human body and facial features, such as MediaPipe<sup>1</sup> (see Figure 1) and OpenPose<sup>2</sup>, we have identified several data types to collect for our dataset. These technologies, developed between 2017 and 2019, provide the opportunity to analyze and validate the unique characteristics of sign communication in different emotional states, as well as for questions or statements. It inspired us to specifically collect sentences with grammatical sentence type marking and marking of emotions to study the role of non-manual in recognition, collecting minimal pairs of signs as potentially challenging for recognition tasks. In the end, we collected quite a wide variety of data types, which are discussed in detail below.

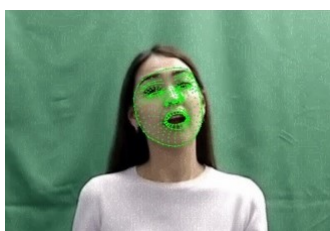


Figure 1: Face landmarks with MediaPipe.

### 3.1. Healthcare videos (2015-2017)

A survey conducted among representatives of the deaf community in Astana and practicing interpreters indicated that deaf signers primarily require accurate interpretations verified by experts for healthcare-related circumstances. Consequently, the initial demand from the community was to establish a comprehensive database for machine learning dedicated to the healthcare domain. All of this involved the development, formation, and collection of a sign language database that encompasses sentences comprising frequently employed medical phrases and terminology.

<sup>1</sup><https://developers.google.com/mediapipe>

<sup>2</sup><https://github.com/CMU-Perceptual-Computing-Lab/openpose>



Figure 2: Frames of healthcare dataset.

Interpreters who have accompanied deaf individuals in medical settings have collaborated to create a list of essential vocabulary terms. The reference interpreter and researchers then constructed sentences to ensure a balanced inclusion of signs in the dataset. Subsequently, we recorded the reference interpreter’s performance of these sign sequences, ensuring that the hands, head, and face remained inside the camera’s field of view and were well-lit. Afterward, we informed the other interpreters that we needed them to replicate his sign sequences since the output videos were for machine-learning algorithms. They agreed to reproduce the sign sequences in full, following the example of the reference interpreter. All 8846 videos were recorded using the website’s tool, which stored them directly on the server. Once the entire dataset had been collected, interpreters were given the task of assessing each other and providing annotations for their colleagues (see Figure 3).

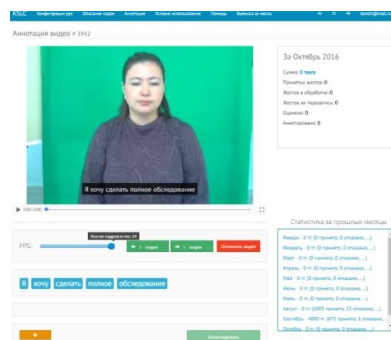


Figure 3: Annotation tool.

We ended up with approximately 148 unique sentences, choosing the top 5 repetitions based on performance quality. Unfortunately, basic CNNs and the Weka tool (Thornton et al., 2013) exhibited a relatively low recognition rate of approximately 53%. The involvement of only four interpreters, three recording modes, and storing videos on the website’s server at 320x240 resolution undoubtedly impacted the output.

### 3.2. Healthcare images (2015-2017)

Revising outputs and drawbacks - we decided to extract images of the most frequent hand configurations to obtain a hand image dataset for training purposes. The idea was to extract cropped images of handshapes (as shown in Figure 4), which will be used for training purposes later.

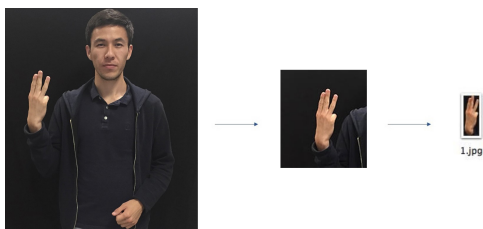


Figure 4: The frame, the ROI, and the element of the dataset.

At first, we decided to try it on a well-known dataset. We downloaded the NCSLGR handshapes videos dataset<sup>3</sup>. We took each 5th frame from videos, which let us obtain hand configurations of various angles. Using a simple hand detector, we extracted configurations by saving ROIs as images - we obtained the set of hand images. Then made the same for our videos.

Next, using HOG (Dalal and Triggs, 2005)+KMeans (MacQueen et al., 1967) clustering, we distributed the same configurations from different subsets to the separate folders for further training (see Figure 5).

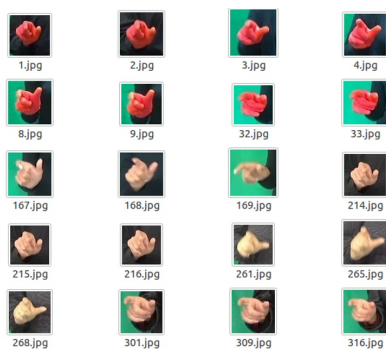


Figure 5: Obtained hand configuration images dataset.

With this technique, we obtained 27 configurations (folders) of the highest inclusion numbers. We implemented a similar HOG+KMeans approach later in Mukushev et al. (2020a) too.

During that period, approaches associated with the generation of supplementary artificial data for training purposes seemed unrealistic. So we made

<sup>3</sup><https://www.bu.edu/asllrp/cslgr/pages/ncslgr-handshapes.html>

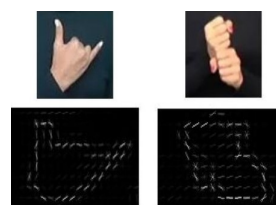


Figure 6: The HOG descriptor performance.

research and tests on various detectors and descriptors available at those time, such as local invariant descriptors: SIFT (Lowe, 1999), SURF (Bay et al., 2006), RootSIFT (Arandjelović and Zisserman, 2012); Binary descriptors: ORB (Rublee et al., 2011), BRISK (Leutenegger et al., 2011), and HOG descriptor. Considering all the advantages and disadvantages of the aforementioned descriptors, we have chosen to utilize the HOG descriptor (see Figure 6) in conjunction with the classification algorithm SVM (Boser et al., 1992) since SVM is reported to exhibit higher performance in cases where there is a lack of data.



Figure 7: Hand configurations from Polish, American and local SL dataset (merged dataset).

We also added images of the same configurations from the Polish SL dataset and got the merged dataset (see Figure 7). After that, we selected 10 configurations with 100 samples and implemented HOG+SVM, results and rates described in Imashev (2017).

### 3.3. Six emotions

The origins of theories regarding fundamental emotions can be traced back to ancient Greece and China as stated by Russell (2003). The fundamental idea of emotions has exerted significant influence for over fifty years. According to the current basic emotion theory, humans have a finite set of emotions that are considered biologically and psychologically “basic” (Wilson-Mendenhall et al., 2013). These emotions exhibit regular recurrence of consistent patterns (Russell, 2006). Researchers in Ekman et al. (2013) revealed evidence of prevalence for six specific emotions: anger, fear, sadness, happiness, surprise, and disgust combined with contempt.

We adhered to the conventional roster of six emotions, except one: five emotions (anger, fear, sadness, happiness, surprise) and “sorry”.



We compiled a list of sentences that are semantically compatible with each of the emotions, in collaboration with K-RSL interpreters. During the recording, the sentences were represented as sequences of glosses via a separate monitor in front of them. Each interpreter performed sentences in different order depending on the emotion. The list of sentences is in Appendix A.



Figure 8: The six emotions in our dataset.

### 3.4. Phonological minimal pairs

Analogous to the existence of phonological minimal pairs in spoken languages, a comparable phenomenon is observed in sign languages (Sandler, 2012; Thompson et al., 2013). In sign language, a minimal pair is a pair of signs with distinct meanings that are distinguished by only one of the major parameters, such as hand configuration, orientation, movement, or non-manual features. Minimal pairs can pose potential problems for recognition tasks, as they are formally similar but semantically different.

There are precedents in the literature for building datasets that specifically target minimal pairs for recognition purposes. As an example, the LIBRAS-UFOP (Cerna et al., 2021). This dataset contains 56 classes of minimal pairs of Brazilian Sign Language. The data was collected using a Microsoft Kinect V1 sensor, which provided comprehensive skeleton data. The dataset was evaluated for recognition using Convolutional Neural Networks (CNN) and long short-term memory (LSTM). The highest accuracy achieved was 74.25%.

The initial reference to phonological minimum pairs in Kazakh-Russian Sign Language was documented in Imashev et al. (2020).

Here are sentences and visual representations for phonological pairs such as RIGHT - MAY (see Table ?? and Figure 9 upper row), and BLUE - WEDNESDAY(v1) (see Table ?? and Figure 9 lower row). Figure 9 also shows two variants for the concept of WEDNESDAY. Note that WEDNESDAY(v1) and WEDNESDAY(v2) are examples of lexical vari-

ability, but only one of them forms a minimal pair with the sign BLUE. This serves as an illustrative example of a case where one sign can be part of a phonological minimal pair and a case of variability simultaneously.



Figure 9: RIGHT(legal) - MAY (upper row), BLUE - WEDNESDAY(v1) - WEDNESDAY(v2) (lower row).

Overall, we collected sentences and videos of 8 pairs and 3 triplets.

### 3.5. Question vs. Statement

Question signs in K-RSL, like question words in spoken/written Kazakh and Russian languages, can be employed not only in interrogative sentences, but also in declarative sentences: “The place **where** sun never sets” and “**Where** are you going?”. Thus, any question sign can occur either with non-manual question marking (eyebrow rise, sideward or backward head tilt) or without it. Furthermore, question marks are accompanied by the mouthing articulation of the related word (see Figure 10).

Question signs are distinguished based on manual aspects, but additional information is obtained through mouthing, which aids recognition. Hence, the two categories of non-manual indicators, namely eyebrow and head position versus mouthing, have distinct functions in recognition. The former aids in distinguishing between statements and questions, while the latter assists in distinguishing between different question signs. To test and confirm, we selected ten question words and constructed twenty phrases: 10 questions and 10 sentences for each word for this dataset (see sentences for WHO in Table ??).

Five interpreters were given them in written form on a screen in front of them one by one to perform (Imashev et al., 2020), the outputs of sign language recognition implementation with this dataset are described in Mukushev et al. (2020b).





Figure 10: A - WHEN, B - WHEN in question; C - HOWMUCH, D - HOWMUCH in question; E - WHERE(location), F - WHERE(location) in question; G - HOW, H - HOW in question; I - WHICH, J - WHICH in question; K - WHATFOR (reason), L - WHATFOR (reason) in question; M - WHICHONE , N - WHICHONE in question, O - WHERE(direction); P - WHERE(direction) in question; Q - WHO, R - WHO in question; S - WHAT/THAT, T - WHAT/THAT in question.

### 3.6. Statements, polar and content questions

For this task, we composed 10 sequences as statements, polar, and wh- questions (see Table ??). We requested interpreters to perform all of them with emotions (in a neutral, surprised, and angry manner) to figure out how emotions and grammatical marking interact in the non-manual features. As mentioned before, deaf communities are quite gated, and this was the first contact and involvement of local native deaf signers in research: several of them (half of the individuals who appeared in this dataset) performed these sentences. Several other deaf signers requested to evaluate and try to recognize emotions (see Figure 11), the results described by Kimmelman et al. (2020). Besides, Kimmelman et al. (2020) is specifically about studying how eyebrow position is affected by sentence type marking and emotions.

### 3.7. K-RSL-173 (Nov. 2019-2020)

After completing a collection of several narrow-purposed subsets, we returned to the idea of collecting a dataset that contains a wide range of concepts used in everyday life. Taking into account the shortcomings of such datasets as PHOENIX (only 9 signers, and a narrow vocabulary about weather and regions of Germany) and DEVISIGN (the participants' performance looked a little unnaturally slow, and the gaze often looked like the



Figure 11: A statement, polar and wh- questions performed in three mood states.

performer did not know the meaning of the signs performed) provide us hints on how to collect our linguistically rich dataset with general, everyday life sentences performed mainly by native signers, fluent signers of different ages, and also filmed in different conditions. By gradually disseminating information about our research, working closely with interpreters for several years, and thereby increasing the level of trust in us from the deaf community, we were able to gather a sufficient number of deaf signers who agreed to participate in data collection and understand the importance for the community.

Initially, we composed 246 sentences, which were revised and narrowed down to 173 sentences with feedback from the reference interpreter, Khasan Israilov. For example, a sentence like 'A doctor told me I needed to remain in bed' (DOCTOR TOLD ME I NEED REMAIN BED REST REGIME), deaf signers will probably perform in a simplified manner as DOCTOR TOLD BED. We recorded these sentences produced by 50 signers (32 deaf, 6 hard of hearing, also 9 hearing CODA, and 3 hearing SODA, including 7 of them are also interpreters).

For sentence translation, we recorded translations of the most proficient (recognized by interpreters and the community) reference interpreter, who made his translations from written sentences, which were composed of spoken language in the manner closest to glosses to avoid any miscommunication. Initially, participants were asked to repeat sign after sign after him from videos. The first few people repeated this but said that they wanted to perform it differently. The next few people were given complete freedom; as a result, the translations of one sentence were completely different from each other (for example: MAY YOU PLEASE SAY TIME vs. just performing sign TIME with ques-

tion face). This led to the fact that we could not collect the required number of sign inclusions for these participants. Therefore, we decided to allow the participants partial freedom with the opportunity to add any clarifications that they consider necessary or change the order of signs.

We detected sign variability at the start of the data collection process mode when participants had partial freedom. After reviewing videos from several initial participants, it was evident that there would be more variability occurrences in the dataset. It presented the opportunity to find specific examples of sign variability in the less explored K-RSL.

It also provided the basis for identifying the variability of signs — one of the reasons for dissatisfaction and arguments like “I do not want to perform signs the same”; there were also formulations like “I used to perform this sign differently”. It helped us identify a certain number of cases of sign variability. See also [Kimmelman et al. \(2022\)](#) for a study on the lexical variability of isolated signs in RSL conducted in partnership with the Garage Museum of Contemporary Art.

Regarding sign variability, consider one of the concepts with several options that was detected in the current dataset. Three configurations used for LEISURE are in Figure 12 also may differ in motions (see Figure 13).

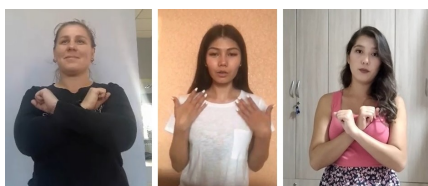


Figure 12: Three variants of **LEISURE** detected in the Dataset.

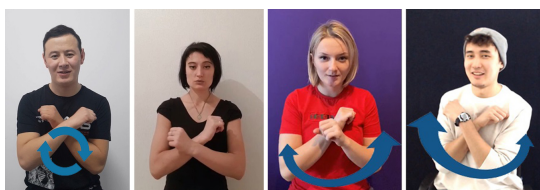


Figure 13: Different motions used for **LEISURE**.

It is noteworthy that all professional interpreters and several native deaf signers performed sign **LEISURE** in the same manner: the hands intersected in the wrist region. The dorsal sides of the clenched fists are in opposition to each other. This configuration rotates in a circular motion in front of the chest (see Figure 14). This observation may indicate the establishment of standardization, at least in the context of interpreting. Alternatively, it

could reveal that these participants share a common geographical or educational background that sets them apart from other signers.

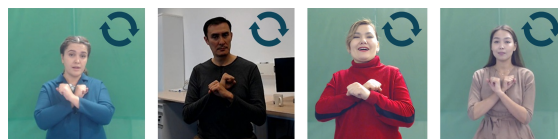


Figure 14: All interpreters performed in the same manner.

Another interesting phenomenon we have observed in the dataset is the presence of polysemic signs, more specifically, those that are distinguished by mouthing. Figure 15 displays different lexical variants of the sign SPOUSE, organized in columns and combined with the mouthing for WIFE or HUSBAND, arranged in rows.

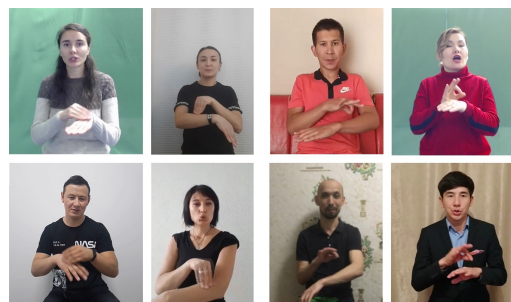


Figure 15: **SPOUSE** variants in handshapes and performance.

An example of a similar phenomenon case is described in [Antonakos et al. \(2015\)](#), German Sign Language Corpus The SIGNUM contains videos for concepts BRUDER and SCHWESTER which utilize the same sign but differ in mouthing (see Figure 16).

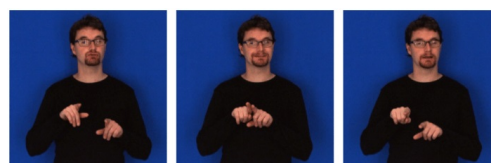


Figure 16: ‘die Geschwister’ sign used for both meanings ‘Bruder’ (brother) and ‘Schwester’ (sister) ([Von Agris et al., 2008](#); [Konrad et al., 2020](#)).

We also discovered two neologisms in the dataset one resulting from the combination of two signs (see Figure 17 a) and the other arising from the combination of two concepts (see Figure 17 b).

In the end, we detected 43 cases of variability (2-6 variants each) and 2 cases of polysemy appearing in the dataset, all of the aforementioned



Figure 17: a) Instagram, b) Facebook.

nuances make it closer to natural sign language performance and more challenging for recognition tasks (Mukushev et al., 2022b).

#### 4. Unpublished Datasets and Future Work

Since deaf individuals often communicate in public settings, the actions of others or external circumstances can disturb the background view. Algorithms that exhibit high accuracy rates under controlled laboratory conditions may perform worse when confronted with unpredictable real-world conditions. Given the difficulty of collecting a dataset in natural environments like parks or public places such as shopping malls, researchers should consider utilizing pre-existing video datasets with uniform backgrounds for keying purposes (see Figure 18). By training algorithms to achieve higher recognition rates in scenarios resembling crowded locations, this approach has the potential to improve sign recognition rates in real-world conditions.

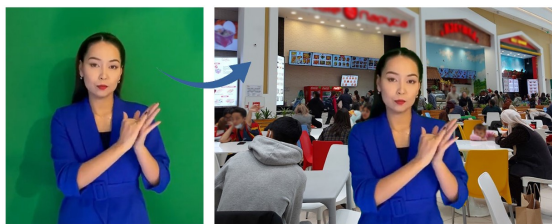


Figure 18: Possible dataset keying.

Priorly acquired datasets can also be utilized as the foundation for generating datasets of 3D signing motion models. For instance, reusing our datasets to get 3D motion files from videos could be expanded to initiate a 3D Signing Dataset (see Figure 19).

Incidentally, amidst the circumstances posed by COVID-19 restrictions, A. Kydyrbekova diligently collected online school lessons aired on National TV, which broadcasted with sign language support



Figure 19: Data-driven signing agent (avatar).

(Mukushev et al., 2022a). Besides, a vocabulary dataset has been collected with 4 interpreters. This dataset contains topics like groceries, household items, also local notions and concepts such as musical instruments, dishes, etc. These two datasets will be available and provided at a later time.

#### 5. Acknowledgements

We are deeply indebted and would like to express our deepest appreciation to the interpreters Gulmira Baizhanova, Khassan Israilov, Aidana Otegenova, Viktoria Antonishina, Samal Nurym, and Shyryn Kozhbanova for their knowledge and support, without which this research would not have been possible. We would also like to extend our deepest gratitude to Dr. Kimmelman for consistently staying connected and offering insights from a sign language perspective. Special thanks to Zhangeldy Bekbatyrov, Qyzylgul Sembina, Shynggys Islam, Azat Kassymgaliyev, Meir, Muslima Karabalayeva, Adai Shomanov, and Aigerim Kydyrbekova. Grateful acknowledgment to Aibek Makazhanov, Zhandos Yessenbayev, and Anara Sandygulova for supporting research on Kazakh-Russian Sign Language throughout the past decade. Research grant awards: 1) The targeted program O.0743 (0115PK02473) of the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (2015-2017); 2) Nazarbayev University Faculty Development Competitive Research Grant Program 2019-2021 “Kazakh Sign Language Automatic Recognition System (KSLARS)”. Award number is 110119FD4545; 3) Nazarbayev University Faculty Development Competitive Research Grant Program 2022-2024, “Kazakh-Russian Sign Language Processing: Data, Tools, and Interaction”; the award number is 11022021FD2902.

#### 6. Bibliographical References

Epameinondas Antonakos, Anastasios Roussos, and Stefanos Zafeiriou. 2015. *A survey on mouth modeling and analysis for sign language recognition*. In *2015 11th IEEE International Conference*



- and *Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–7. IEEE.
- Relja Arandjelović and Andrew Zisserman. 2012. [Three things everyone should know to improve object retrieval](#). In *2012 IEEE conference on computer vision and pattern recognition*, pages 2911–2918. IEEE.
- Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Quan Yuan, and Ashwin Thangali. 2008. [The american sign language lexicon video dataset](#). *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.
- ALC Barczak, NH Reyes, M Abastillas, A Piccio, and Teo Susnjak. 2011. [A new 2d static hand gesture colour image dataset for asl gestures](#).
- Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. [Surf: Speeded up robust features](#). In *Computer Vision—ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer.
- Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. [A training algorithm for optimal margin classifiers](#). In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.
- Lourdes Ramirez Cerna, Edwin Escobedo Cardenas, Dayse Garcia Miranda, David Menotti, and Guillermo Camara-Chavez. 2021. [A multimodal libras-ufop brazilian sign language dataset of minimal pairs using a microsoft kinect sensor](#). *Expert Systems with Applications*, 167:114179.
- Xiujuan Chai, Hanjie Wang, and Xilin Chen. 2014. The design of large vocabulary of chinese sign language database and baseline evaluations. In *Technical report VIPL-TR-14-SLR-001. Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS)*. Institute of Computing Technology.
- Navneet Dalal and Bill Triggs. 2005. [Histograms of oriented gradients for human detection](#). In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.
- Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. 2013. [Emotion in the human face: Guidelines for research and an integration of findings](#), volume 11. Elsevier.
- Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. 2012. [Rwth-phoenix-weather: A large vocabulary sign language recognition and translation corpus](#). In *LREC*, volume 9, pages 3785–3789.
- Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. [Video-based sign language recognition without temporal segmentation](#). In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pages 2257–2264. AAAI press.
- Alfarabi Imashev. 2017. [Sign language static gestures recognition tool prototype](#). In *2017 IEEE 11th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.
- Alfarabi Imashev, Medet Mukushev, Vadim Kimmelman, and Anara Sandygulova. 2020. [A dataset for linguistic understanding, visual evaluation, and recognition of sign languages: The k-rsl](#). In *Conference on Computational Natural Language Learning*.
- Michal Kawulok, Tomasz Grzejszczak, Jakub Nalepa, and Mateusz Knyc. 2013. [Database for hand gesture recognition](#).
- Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. [Eyebrow position in grammatical and emotional expressions in kazakh-russian sign language: A quantitative study](#). *PLoS one*, 15(6):e0233731.
- Vadim Kimmelman, Anna Komarova, Lyudmila Luchkova, Valeria Vinogradova, and Oksana Alekseeva. 2022. [Exploring networks of lexical variation in russian sign language](#). *Frontiers in psychology*, 12:740734.
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. [Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release](#).
- P Pramod Kumar, Prahlad Vadakkepat, and Ai Poh Loh. 2010. [Hand posture and face recognition using a fuzzy-rough approach](#). *International Journal of Humanoid Robotics*, 7(03):331–356.
- Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. 2011. [Brisk: Binary robust invariant scalable keypoints](#). In *2011 International conference on computer vision*, pages 2548–2555. IEEE.



- David G Lowe. 1999. [Object recognition from local scale-invariant features](#). In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. IEEE.
- James MacQueen et al. 1967. [Some methods for classification and analysis of multivariate observations](#). In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.
- A.M. Martinez, R.B. Wilbur, R. Shay, and A.C. Kak. 2002. [Purdue rvl-slll asl database for automatic recognition of american sign language](#). In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 167–172.
- Medet Mukushev, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. 2020a. [Automatic classification of handshapes in russian sign language](#). In *Proceedings of the LREC 2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. European Language Resources Association (ELRA).
- Medet Mukushev, Aigerim Kydyrbekova, Vadim Kimmelman, and Anara Sandygulova. 2022a. [Towards large vocabulary Kazakh-Russian Sign Language dataset: KRSL-OnlineSchool](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 154–158, Marseille, France. European Language Resources Association.
- Medet Mukushev, Arman Sabyrov, Alfarabi Imashev, Kenessary Koishibay, Vadim Kimmelman, and Anara Sandygulova. 2020b. [Evaluation of manual and non-manual components for sign language recognition](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).
- Medet Mukushev, Aidyn Ubingazhibov, Aigerim Kydyrbekova, Alfarabi Imashev, Vadim Kimmelman, and Anara Sandygulova. 2022b. [Fluentsigners-50: A signer independent benchmark dataset for sign language processing](#). *PLoS ONE*, 17.
- Carol Neidle, Ashwin Thangali, and Stan Sclaroff. 2012. [Challenges in development of the american sign language lexicon video dataset \(asllvd\) corpus](#). In *5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC*.
- Rie Nishio, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Thomas Hanke, and Christian Rathmann. 2010. [Elicitation methods in the dgs \(german sign language\) corpus project](#). In *sign-lang@ LREC 2010*, pages 178–185. European Language Resources Association (ELRA).
- Oğulcan Özdemir, Ahmet Alp Kindiroğlu, Necati Cihan Camgoz, and Lale Akarun. 2020. [BosphorusSign22k Sign Language Recognition Dataset](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*.
- Pramod Kumar Pisharady, Prahlad Vadakkepat, and Ai Poh Loh. 2013. [Attention based detection and recognition of hand postures against complex backgrounds](#). *International Journal of Computer Vision*, 101:403–419.
- Nicolas Pugeault and Richard Bowden. 2011. [Spelling it out: Real-time asl fingerspelling recognition](#). In *2011 IEEE International conference on computer vision workshops (ICCV workshops)*, pages 1114–1119. IEEE.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. [Orb: An efficient alternative to sift or surf](#). In *2011 International conference on computer vision*, pages 2564–2571. IEEE.
- James A Russell. 2003. [Core affect and the psychological construction of emotion](#). *Psychological review*, 110(1):145.
- James A Russell. 2006. [Emotions are not modules](#). *Canadian Journal of Philosophy Supplementary Volume*, 32:53–71.
- Wendy Sandler. 2012. [The phonological organization of sign languages](#). *Language and linguistics compass*, 6(3):162–182.
- Robin L Thompson, David P Vinson, Neil Fox, and Gabriella Vigliocco. 2013. [Is lexical access driven by temporal order or perceptual salience? evidence from british sign language](#). In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.
- Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. [Auto-weka: Combined selection and hyperparameter optimization of classification algorithms](#). In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 847–855.

Jochen Triesch and Christoph Von Der Malsburg. 1996. [Robust classification of hand postures against complex backgrounds](#). In *Proceedings of the second international conference on automatic face and gesture recognition*, pages 170–175. IEEE.

Jochen Triesch and Christoph Von Der Malsburg. 2001. [A system for person-independent hand posture recognition against complex backgrounds](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1449–1453.

Ulrich Von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. [The significance of facial features for automatic sign language recognition](#). In *2008 8th IEEE international conference on automatic face & gesture recognition*, pages 1–6. IEEE.

Christine D Wilson-Mendenhall, Lisa Feldman Barrett, and Lawrence W Barsalou. 2013. [Neural evidence that human emotions share core affective properties](#). *Psychological science*, 24(6):947–956.

## 7. Appendix A. Sentences composed for six emotions dataset

Table 3: Sentences on 6 selected emotions

<b>Anger</b>	<b>Sadness</b>
People's anger	My memories of the past are sad
There is no need to rush - you will become angry	Sad face
Patience, you do not need to be angry	Sad eyes
Anger - is a strong feeling	They are sad
Anger prevents thinking rationally	I hear his voice is sad
Strong anger	There is no need to be sad
Anger helps to win	Sadness ends soon
When he is angry, everyone is scared	Happy and sad
Old people are angry	Looked away with a sad look
They are angry for no reason	Why are you sad
<b>Fear</b>	<b>Surprised</b>
Fear of the dark	Childhood is when everything is surprising
People struggle with their fears	Their knowledge is surprising
Fear is hard to hide	Are you surprised?
We are afraid of many things	Kazakhstan's nature is surprisingly beautiful
There is no need to be scared	The boy looked surprised
Fear has big eyes	Fairytales are surprising
Fear helps the enemy	The athletes' records are surprising
Very scary movie	Surprised faces
Grandmother fears the future	These discoveries are surprising for us
She was afraid of heights	They looked into the distance in surprise
<b>Sorry</b>	<b>Happy</b>
I'm sorry, and I'm suffering	Well-being is the source of happiness
You are always feel sorry	Serene happiness
Being able to be sorry is important for the future	True happiness
I feel sorry for him; that's why crying	I'm happy
Grandma always feels sorry for everyone	This is the reason for happiness
People must be kind and be able to feel sorry for each other- otherwise, the world has no future	Happy face
I'm sorry for the thrown-away books	A happy man
I'm really sorry	I found a job - I'm happy
I feel sorry for the animals	They are happy that they came
I'm sorry - I left	We are happy that we left

# Enhancing Syllabic Component Classification in Japanese Sign Language by Pre-training on Non-Japanese Sign Language Data

Jundai Inoue, Makoto Miwa , Yutaka Sasaki , Daisuke Hara

Toyota Technological Institute

2-12-1 Hisakata, Tempaku-ku, Nagoya, Aichi, 468-8511, Japan

{sd24410, makoto-miwa, yutaka.sasaki, daisuke}@toyota-ti.ac.jp

## Abstract

In sign languages, syllables are composed of syllabic components consisting of locations, movements, and handshapes; however, the rules of combinations of these syllabic components are still unclear. Decomposing existing syllables into syllabic components is necessary to clarify the rules. This study aims to construct an automatic syllabic component classification system for Japanese Sign Language (JSL) using deep learning. We propose a pre-training method using non-Japanese Sign Language data to achieve high performance in classifying syllabic components in a situation where the number of training JSL videos is limited. We also investigate multitask learning for syllabic component classification to share the information among the syllabic components. Experiments on the syllabic component classification for the dominant hand show that 1) pre-training with the American Sign Language (ASL) dataset improved classification performance for the movement and handshape components and 2) multitask learning did not contribute to the performance improvement of syllabic component classification. We also investigated the influence of pre-training on syllabic component classification by visualizing critical elements in videos to predict the components.

**Keywords:** Japanese Sign Language, Syllabic components, Pre-training, Multitask learning

## 1. Introduction

*Locations, movements, and handshapes* are the syllabic components in sign languages. Syllables of sign language are combinations of the syllabic components, and the composition rules for the syllables are still unclear (Hara, 2016). To analyze the rules of syllable composition in Japanese Sign Language (JSL), Hara (2019) proposed a syllable database with videos of syllables and their components that are decomposed by hand. However, manually decomposing a number of syllables that have not yet been registered in the database into syllabic components is costly. Therefore, it is needed to construct a system that can automatically recognize syllabic components from JSL videos. The syllabic component recognizer could be used not only to supplement the database but also to further analyze JSL using the system's prediction results.

Recently, deep learning approaches to sign language processing have been shown to be effective (Jiang et al., 2021; Chen et al., 2022; Zuo et al., 2023). Deep learning methods require a large amount of labeled training data to achieve high performance, but unfortunately, the number of JSL videos with labeled syllabic components is limited. On the other hand, there is a large amount of data of a non-Japanese Sign Language, such as American Sign Language (ASL), and the two sign languages share features in expressing signs with manual and non-manual signals. Although we can expect the improvement of classification perfor-

mance for JSL by using the shared features, such an approach has yet to be investigated.

This study aims to construct an automatic syllabic component classification system from JSL videos. As the first step toward this goal, this study focuses on the location, movements, and handshape of the dominant hand. To address the problem of limited data in JSL, we propose pre-training using non-JSL datasets. We conduct training on JSL video data to classify syllabic components after initializing the parameters with those trained on a non-JSL dataset. We also introduce multitask learning in classifying location, movement, and handshape components by sharing the base classification model among the components.

The contributions of this study are summarized as follows:

- We constructed a system that automatically recognizes syllabic components of the dominant hand from JSL videos.
- We showed the effectiveness of using models pre-trained on a non-JSL dataset for the movement and handshape classification from JSL with limited data.
- We found that information sharing between tasks does not necessarily improve classification performance through multitask learning of syllabic components in JSL.



## 2. Related Work

### 2.1. Japanese Sign Language Dataset

Nagashima et al. (2018) constructed a versatile JSL database that can be used in the fields of linguistics and engineering. The database includes high-resolution video data capturing the actions of two native signers with a high-resolution camera from the front and diagonally forward from the left and right. Additionally, it incorporates 3D motion data obtained through optical motion capture and depth data from distance sensors. The dataset provides data on 4,873 glosses and ten dialogues.

Hara (2019) defined a JSL coding manual and created a syllable database in which the syllables were broken down into location, movement, and handshape components. The database contains video clips representing the JSL syllables, recorded with a single signer. 1,086 syllable videos were included, each consisting of approximately 300 frames. The location components are classified into 22 categories to indicate the hand locations in space or on the body. The handshape components are divided into 69 categories. The location and handshape components are assigned to a single category label in the video. The location component signifies the starting position of the sign, and the handshape component indicates the shape of the hand. We should note that this database manually defines base handshapes so that each syllable can be represented by a single base handshape. We use this base handshape as the handshape component, and the changes in the handshape are represented by the movement component.

The movement components are distinguished into 55 ways of moving a hand, such as rightward movement and finger joint opening, with one to three categories assigned to each video. In addition to the components for dominant and non-dominant hands, more detailed decompositions of each syllabic component are attached, such as “contact,” “hand orientation,” and “metacarpal orientation.”

### 2.2. Sign language processing using machine learning and deep learning

Sign language processing using machine learning and deep learning, such as Sign Language Recognition (SLR) for predicting gloss (Jiang et al., 2021; Zuo et al., 2023) and sign language translation for translating signs into spoken language (Chen et al., 2022), has been actively conducted. Skeleton Aware Multi-modal SLR (SAM-SLR) (Jiang et al., 2021) is a framework that integrates body, motion, and depth information in addition to video and keypoint information. Video-Keypoint Network (VKNet) (Zuo et al., 2023) extracts features from 64 and 32 video frames and keypoints to account for

different temporal information. VKNet consists of two sub-networks, VKNet-64 and VKNet-32. Each sub-network also contains video and keypoint encoders, and there are bidirectional lateral connections (Duan et al., 2022) to exchange information between each encoder. S3D (Xie et al., 2018), a 3D Convolutional Neural Network that can consider spatio-temporal information, is used as the encoder. After keypoints are estimated from the video using a learned pose estimation model, HR-Net (Sun et al., 2019), 64 and 32 video frames and keypoints are input to VKNet-64 and VKNet-32, respectively. The combined representation vectors from each network are used to predict the gloss. VKNet performed well on several datasets for SLR.

Studies on sign languages considering syllabic components have also been conducted (Zhang and Duh, 2023; Tavella et al., 2022; Kezar et al., 2023; Hatano et al., 2016). To clarify the importance of the handshape component in SLR, Zhang and Duh (2023) constructed a dataset labeled with handshapes on an existing SLR dataset and proposed a model that predicts both glosses and handshapes simultaneously by extending the existing SLR model. The proposed model performs better than those that only use videos as input without considering handshapes. Tavella et al. (2022) and Kezar et al. (2023) have constructed datasets labeling multiple syllabic components in addition to gloss in sign language videos. Furthermore, Kezar et al. (2023) classified 16 different phonological features, which are close to fine-grained syllabic components, and demonstrated that learning the features through classification contributes to improving the performance of SLR. In JSL, Hatano et al. (2016) employed machine learning methods to recognize the location, movement, and handshape components and construct a SLR system based on the weighted sum of classification scores for each component. This method requires extracting the video’s features, such as coordinates, velocity, and acceleration.

## 3. Methods

This study proposes a method for classifying syllabic components in JSL videos using pre-training on a non-JSL dataset. This study focuses on the location, movement, and handshape components of the dominant hand, which are defined in the syllable database created by Hara (2019) and employs VKNet (Zuo et al., 2023) as the base deep learning model. We initialized the parameters of VKNet with those pre-trained on a non-JSL dataset to leverage information from non-JSL. The overall architecture of the proposed model is illustrated in Figure 1.

As explained in §2.1, there are 22, 55, and 69 categories for location, movement, and handshape

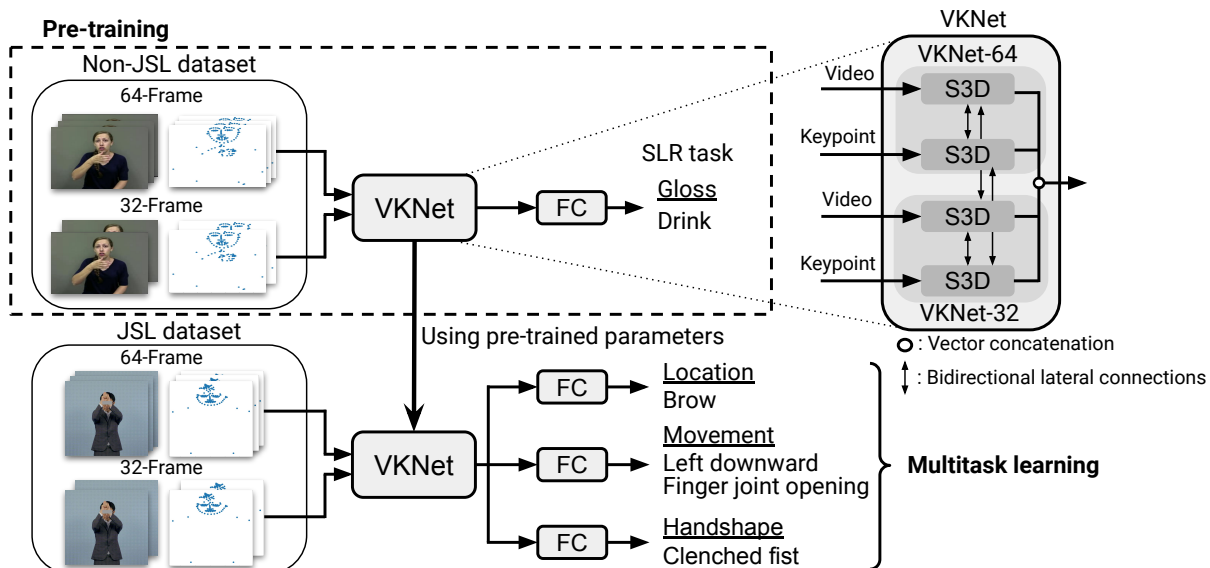


Figure 1: The overview of syllabic component classification through pre-training using non-JSL dataset

components, respectively. We added three fully connected (FC) layers corresponding to individual components to the VKNet pre-trained on the non-JSL to classify each syllabic component.

A softmax function is applied to the output vector of the FC layers for the location and handshape components, where a single label is assigned from multiple categories. This function enables multi-class classification, where the class with the highest predicted probability is considered the prediction. By contrast, a sigmoid function is applied to the output vector of the FC layers for the movement component, which involves multiple labeled movements. This function allows for binary classification for each movement type; movements with predicted probabilities higher than a threshold are considered the prediction in the multi-label classification.

The loss function includes cross-entropy and asymmetric losses (Ridnik et al., 2021). The cross-entropy loss is used for location and handshape classification, while the Asymmetric Loss (AsLoss) is applied to the movement classification. Since there are only up to three movements for each syllable in the database, the classification problem is highly imbalanced, with few positive and many negative examples. The AsLoss addresses this imbalance by calculating a weighted sum in which the weight of the loss in positive examples is larger than that in negative examples. It is defined as:

$$\text{AsLoss} = \begin{cases} -(1-p)^{\gamma^+} \log(p) & \text{if } y = 1 \\ -p_m^{\gamma^-} \log(1-p_m) & \text{otherwise} \end{cases} \quad (1)$$

where  $p_m$  is defined in Equation (2) to ignore negative examples that can be classified easily.

$$p_m = \max(p - m, 0) \quad (2)$$

Note that  $p$  is the network’s output probability and hyperparameters  $\gamma^-$  and  $\gamma^+$  are sets such that  $\gamma^- > \gamma^+$  to emphasize the contribution of positive examples.  $m$  represents the threshold value.

During training, multitask learning is performed to share the information among syllabic components. Specifically, VKNet is shared, and the loss function is the sum of classification losses for each syllabic component.

## 4. Experimental settings

We evaluated the proposed method using the syllable database created by Hara (2019). We randomly split the 1,072 instances annotated with the location, movement, and handshape components into 750, 161, and 161 instances for training, development, and testing, respectively. The statistics for the top-10 instances of each component are presented in Table 1. The table shows that syllable instances are highly imbalanced among the categories. To avoid highly challenging classification problems, we excluded instances with the categories with fewer than five instances in the training data, treating them as false-negative predictions. We adopted the micro F-score as the evaluation metric.

As the pre-training parameters, we utilized the pre-trained VKNet parameters,<sup>1</sup> which was trained on the 14,289 training instances with 2,000 glosses of Word-Level American Sign Language (WLASL) dataset for SLR in ASL (Li et al., 2020).

We conducted two comparisons in the experiments. The first comparison is to investigate the ef-

<sup>1</sup><https://github.com/FangyunWei/SLRT/tree/main/NLA-SLR>













Movement	#	Handshape	#	Location	#
Rightward movement of a hand	142		138	 *	835
Forward movement of a hand	135		125	Temples	40
Wrist rotation: outward rotation of a wrist with the little finger as the axis	120		57	Mouth	32
Downward movement of a hand	117		55	Chest	23
Flexion of finger joints with handshape changes	80		53	Brow	22
Extension of finger joints with handshape changes	77		48	Eyes	17
Circular or semicircular movement on a horizontal plane	69		42	Face	16
Upward movement of a hand	64		40	Elbow	13
Leftward movement of a hand	61		40	 **	13
Non-linear movement (trajectory) of a hand	51		34	Abdomen	12

Table 1: Numbers (#) of top-10 instances for the location, movement, handshape components, icons from McKee et al. (2011). \* and \*\* in the location component represent the neutral space in which the sign is made in front of the body or face, respectively.

Method	Syllabic component		
	Location	Movement	Handshape
VKNet	80.75 ( $\pm 1.02$ )	38.29 ( $\pm 2.54$ )	39.54 ( $\pm 1.05$ )
+ Pre-training	81.16 ( $\pm 2.05$ )	52.41* ( $\pm 0.86$ )	44.72* ( $\pm 3.55$ )
+ Multitask learning	81.99 ( $\pm 0.00$ )	45.76* ( $\pm 0.82$ )	42.23 <sup>†</sup> ( $\pm 1.34$ )

Table 2: Results of syllabic component classification. The means of three runs are shown as the final micro F-scores (%). The numbers in parentheses are standard deviations. \* and <sup>†</sup> denote significance levels of 0.05 and 0.1 compared with the results directly above.

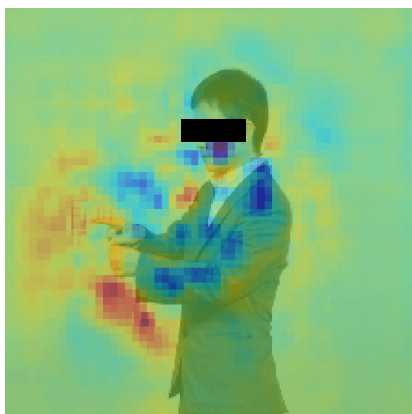
fectiveness of pre-training using the ASL dataset in syllabic component classification for JSL; we compared the classification performance of VKNet with parameters initialized from the pre-trained model and VKNet with randomly initialized parameters. The second comparison is to evaluate multitask learning. We compared the classification performance when simultaneously or independently addressing each task to understand the impact of information sharing between tasks. We used the Adam optimization method (Kingma and Ba, 2015), setting the learning rate to  $5 \times 10^{-5}$  and applied cosine annealing as a scheduler to change the learning rate per epoch. We set the hyperparameters  $\gamma^-$ ,  $\gamma^+$ , and  $m$  of the AsLoss to 4, 1, and 0.05, respectively. To suppress overfitting, we employed dropout (Srivastava et al., 2014) and regularization, setting their values to 0.2 and  $10^{-3}$ , respectively.

## 5. Results

The results of syllabic component classification from JSL videos in test data are shown in Table 2. The results of syllabic component classification

using VKNet with parameters pre-trained on the WLASL dataset as initial values showed that the micro F-scores for the location, movement, and handshape components were improved compared to those using VKNet with random parameters as initial values. The results evaluated on the development and test data are summarized in appendix B. We conducted a significance difference test with the bootstrap method to verify the improvement in classification performance of the pre-trained VKNet. As a result, we confirm that the pre-training method effectively improved the classification of the movement and handshape components of JSL.

Multitask learning improved the micro F-score of the location component but decreased those of the movement and handshape components. The significance test showed a significant decrease in the classification of the movement component, while there was no significant difference for the location and handshape components. This result indicates that multitask learning is ineffective or harmful in classifying syllabic components of JSL.



(a) Visualization result of VKNet's prediction basis



(b) Visualization result of pre-trained VKNet's prediction basis

Figure 2: Visualization results (classification of the movement component)

## 6. Discussion

To verify the influence of pre-training on the syllabic components of VKNet, we visualized the parts of the video VKNet focused on while predicting syllabic components using Adaptive Occlusion Sensitivity Analysis (AOSA) (Uchiyama et al., 2023), one of the methods of explainable AI techniques. The AOSA results were visualized with colors from red to blue to indicate their importance; the areas with high importance are shown in red. The example of the movement component that could not be classified by VKNet but could be classified by the pre-trained VKNet is visualized in Figure 2. From these results, we can see that the right hand, which is the dominant hand, is more focused after pre-training. This change in the focus suggests that the pre-trained VKNet can make more accurate predictions than the VKNet by focusing on the dominant hand and classifying syllabic components.

## 7. Conclusions

This study proposed the classification of the syllabic component for the dominant hand using parameters of a model pre-trained on a non-JSL dataset as a first step to construct a method for syllabic component classification based on JSL videos. We also introduced multitask learning for sharing information among syllabic component classification. We evaluated the proposed method based on the VKNet model using the JSL database in the experiments. Experimental results show that pre-training with the ASL dataset significantly improves the classification performance of the movement and hand-shape components from a limited number of the JSL videos. On the other hand, the classification performance with multitask learning did not improve the performance of syllabic component classification in JSL. We also investigated the effect of pre-training on syllabic component prediction by visualizing the predictive basis of VKNet using AOSA. The visualization results suggest that the proposed pre-training enabled the focus on the target hand. Future work includes investigating the models and training methods to improve the classification and classification performance of syllabic components for both the dominant and non-dominant hands.

## 8. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23H00626.

## 9. References

- Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130.
- Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2969–2978.
- Daisuke Hara. 2016. *18. An information-based approach to the syllable formation of Japanese Sign Language*, pages 457–482. De Gruyter Mouton, Berlin, Boston.
- Daisuke Hara. 2019. *New Japanese Sign Language Coding Manual*. (In Japanese).
- Mika Hatano, Shinji Sako, and Tadashi Kitamura. 2016. *Real-time sign language recognition by*



- kinect v2 based on three elements of sign language. *IEICE technical report*, 115(491):59–64.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Lee Kezar, Jesse Thomason, Naomi Caselli, Zed Sehyr, and Elana Pontecorvo. 2023. [The semlex benchmark: Modeling asl signs and their phonemes](#). In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '23*, New York, NY, USA. Association for Computing Machinery.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- D. McKee, R. McKee, S. Pivac Alexander, L. Pivac, and M Vale. 2011. Online dictionary of new zealand sign language. <https://www.nzsl.nz/>.
- Yuji Nagashima, Daisuke Hara, Yasuo Horiuchi, Shinji Sako, Rituko Kikusawa, Akira Ichikawa, Keiko Watanabe, and Naoto Kato. 2018. Development of the super high-definition and high-precision japanese sign language database available for various research fields. In *Proceedings of Language Resources Workshop*, volume 3, pages 148–155. National Institute for Japanese Language and Linguistics. (In Japanese).
- Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: A simple way to prevent neural networks from overfitting](#). *Journal of Machine Learning Research*, 15(56):1929–1958.
- Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Federico Tavella, Viktor Schlegel, Marta Romeo, Aphrodite Galata, and Angelo Cangelosi. 2022. [WLASL-LEX: a dataset for recognising phonological properties in American Sign Language](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 453–463, Dublin, Ireland. Association for Computational Linguistics.
- Tomoki Uchiyama, Naoya Sogi, Koichiro Niinuma, and Kazuhiro Fukui. 2023. Visually explaining 3d-cnn predictions for video classification with an adaptive occlusion sensitivity analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1513–1522.
- Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.
- Xuan Zhang and Kevin Duh. 2023. [Handshape-aware sign language recognition: Extended datasets and exploration of handshape-inclusive methods](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2993–3002, Singapore. Association for Computational Linguistics.
- Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14890–14900.

## A. Impact of data imbalance on location component classification

For the location component classification, neutral space instances, the first row in Table 1, cover most of the dataset. To examine its impact on the classification results, we used a pre-trained VKNet and evaluated it by excluding the instances. The evaluation results on the development data are shown in Table 3. When excluding neutral space instances from the dataset, the performance significantly dropped. This result suggests that the model was affected by the bias in the dataset and fitted to the neutral space class. This performance degradation indicates that, to improve the generality of the model, the bias in the dataset needs to be addressed by sampling data or changing the loss function.

## B. Overall result

In this study, we set four learning conditions to compare the effects of pertaining VKNet and multitask learning: (1) no pertaining VKNet, no multitask learning, (2) pertaining VKNet, no multitask learning, (3) no pertaining VKNet, multitask learning, (4) pertaining VKNet and multitask learning. We performed syllabic component classification for each condition using the development and test data. The results are shown in Table Table 4

## C. Hyperparameter tuning in multitask learning

we conducted additional experiments to optimize the coefficients of the loss functions for each task in multitask learning. Previously, we summed the losses for each syllabic component. Still, this time, we introduced weighting coefficients for the loss of each syllabic component and attempted to optimize these coefficient values using a Bayesian optimization. Specifically, the value of each coefficient was constrained to be between 0 and 1, and the sum of all coefficients was always set to 1. We performed 70 iterations of Bayesian optimization and searched for the combination of coefficients that maximized the micro F-score for syllabic component classification on the development data. It is shown in Table 5, where the optimal coefficient values obtained by Bayesian optimization and the corresponding micro F-scores are shown in contrast to the micro F-scores obtained by simply summing the losses. After three evaluations, the micro F-score for the handshape component showed a slight improvement, although the micro F-scores for the location and movement components showed a slight decrease. However, these score changes

	Location
pre-trained VKNet w/ neutral space	80.75 ( $\pm$ 1.02)
pre-trained VKNet w/o neutral space	41.67 ( $\pm$ 4.54)

Table 3: Results of location component classification with and without neutral space instances. Neutral space instances constitute a large portion of the dataset. The performance is measured using the micro F-score (%), with the reported values showing the average and standard deviation over three evaluation runs.

were within the margin of error, indicating no significant difference resulted from simply summing the losses for each syllabic component. Therefore, we evaluated the test data using a simple sum of losses with equal weights.

Method		Dev			Test		
		Location	Movement	Handshape	Location	Movement	Handshape
Multitask	VKNet	82.40 ( $\pm$ 1.27)	34.06 ( $\pm$ 0.52)	39.75 ( $\pm$ 1.83)	80.33 ( $\pm$ 1.06)	38.55 ( $\pm$ 1.25)	35.20 ( $\pm$ 2.55)
	+ Pre-training	82.20 ( $\pm$ 1.17)	39.94 ( $\pm$ 2.85)	47.41 ( $\pm$ 2.29)	81.99 ( $\pm$ 0.00)	45.76 ( $\pm$ 0.82)	42.23 ( $\pm$ 1.34)
Singletask	VKNet	83.85 ( $\pm$ 1.01)	34.57 ( $\pm$ 0.39)	43.89 ( $\pm$ 0.29)	80.75 ( $\pm$ 1.02)	38.29 ( $\pm$ 2.54)	39.54 ( $\pm$ 1.05)
	+ Pre-training	83.44 ( $\pm$ 0.77)	44.98 ( $\pm$ 1.06)	47.82 ( $\pm$ 1.02)	81.16 ( $\pm$ 2.05)	52.41 ( $\pm$ 0.86)	44.72 ( $\pm$ 3.55)

Table 4: Results of syllabic component classification with and without pertaining and with and without multitask learning. The evaluation metric is the micro F-score (%). The mean and standard deviation of the three evaluations are shown.

hyperparameter	Dev		
	Location	Movement	Handshape
alpha = 0.095704 beta = 0.597839 gamma = 0.306457	78.46 ( $\pm$ 1.17)	38.70 ( $\pm$ 1.86)	48.24 ( $\pm$ 1.63)
alpha = beta = gamma	82.20 ( $\pm$ 1.17)	39.94 ( $\pm$ 2.85)	47.41 ( $\pm$ 2.29)

Table 5: Micro F-score (%) of syllabic component classification using the optimized hyperparameters obtained from Bayesian optimization and an equal weight baseline. Coefficients for location, movement, and handshape are denoted as alpha, beta, and gamma, respectively.

# Building Your Query Step by Step: A Query Wizard for the MY DGS – ANNIS Portal of the DGS Corpus

Amy Isard 

Institute of German Sign Language and Communication of the Deaf  
and House of Computing and Data Science  
University of Hamburg, Germany  
amy.isard@uni-hamburg.de

## Abstract

*MY DGS – ANNIS* makes the Public DGS Corpus available through the corpus query and visualization tool ANNIS. Due to the complex nature of the corpus, composing queries for advanced research questions can quickly become increasingly complicated. We present a Query Wizard which assists users in building valid queries for *MY DGS – ANNIS*. Complex queries are built up from smaller blocks, which can be linked to each other through context-sensitive connections. Blocks provide options specific to a given annotation tier and dynamically lead users through their construction while preventing the creation of invalid queries. Once completed, queries can be opened directly in *MY DGS – ANNIS*.

**Keywords:** German Sign Language (DGS), corpus query tool, ANNIS, query wizard

## 1. Introduction

In 2022 the DGS-Korpus project introduced *MY DGS – ANNIS* (Isard and Konrad, 2022), a third portal to provide access to release 3 of the Public DGS Corpus (Hanke et al., 2020). ANNIS (Krause and Zeldes, 2016) is a corpus query and visualization tool which allows corpus queries written in the ANNIS Query Language AQL<sup>1</sup> to be performed over multiple annotation tiers and corpus metadata. Our interface allows researchers to search either the German or the English version of the Public DGS Corpus.

*MY DGS – ANNIS* has enabled sign language researchers to make complicated queries over the Public DGS Corpus, but the combination of multiple annotation and metadata tiers with complex glossing conventions on the one hand, and AQL syntax on the other, means that novice users have not always found it easy to create valid queries which exactly match what they were searching for. The ANNIS interface contains a general-purpose Query Builder tool, but the dyadic and sign-based nature of our data (see Section 2.1) necessitates a more customised approach.

Several corpus projects which make their data available through ANNIS have provided “simple search” interfaces for their ANNIS instances (Dipper, 2015), including the Reference Corpus Middle Low German/Low Rhenish (1200–1650)<sup>2</sup> and the Reference Corpus of Middle High Ger-

man (1050–1350)<sup>3</sup>. Inspired by these we decided to create a “simple search” interface for the Public DGS Corpus: the *MY DGS – ANNIS* Query Wizard.

Our Query Wizard allows users to create complex queries out of smaller building blocks by creating connections between them, and uses visual elements to make the connections between the blocks and the resulting AQL query easier to understand. We hope that this will help users to learn about the structure of AQL queries, so that if their needs surpass the scope of the Query Wizard, they will be ready to manually refine queries in *MY DGS – ANNIS*.

The Query Wizard ensures that only valid queries are generated and makes query building easier in a number of ways:

- Users select annotation and metadata tiers from a comprehensive list, ensuring only valid tiers are involved and avoiding issues like spelling errors.
- Instead of writing complex regular expressions to refine search to only certain tokens, users can compose these expressions using context-sensitive check boxes.
- Connections within tiers can be added without knowledge of the exact syntax necessary.

This article introduces the Query Wizard and explains how it integrates with *MY DGS – ANNIS*. In Section 2 we describe the Public DGS Corpus data available through *MY DGS – ANNIS*, with the annotations and metadata in Section 2.1, the ANNIS Query Language (AQL) in Section 2.2 and the *MY*

<sup>1</sup><http://korpling.github.io/ANNIS/4.11/user-guide/aql/index.html>

<sup>2</sup><https://www.slm.uni-hamburg.de/en/ren/korpus/datenzugang/einfache-suche.html>

<sup>3</sup><https://www.linguistics.rub.de/rem/access/simplesearch.en.html>



DGS – ANNIS interface in Section 2.3. In Section 3 we describe the interface and usage of the Query Wizard, and in Section 4 we show how some example queries can be built up. Section 5 contains conclusions and describes further features which we intend to add to the Query Wizard in future.

## 2. MY DGS – ANNIS

### 2.1. Annotations and Metadata

MY DGS – ANNIS provides datasets representing both release 3 and release 4 of the Public DGS Corpus, with two versions of each dataset, one each for the English and German versions of its annotations<sup>4</sup>.

Table 1 lists the main annotation tiers with a brief description of their content. Each element in a tier contains text which can be searched; in the case of translations and mouthings the text is fairly simple, and the HamNoSys tier can be searched by inputting HamNoSys characters directly, which can be done using the HamNoSys editor.<sup>5</sup>

For Gloss and GlossType tiers a special syntax is used which makes search more complex. In these tiers, each token is represented by a type gloss. Each type gloss contains a gloss word, one or two digits which denote different lexical variants, and an optional letter denoting phonological variants. Types that denote form without specifying meaning (i.e. they are supertypes rather than subtypes) are indicated by the caret character (^). In the Gloss tier, an asterisk (\*) indicates that a token gloss diverges in some way from the citation form of the type<sup>6</sup>. We provide one gloss tier for each participant to enable collocation searches within a tier (for an example see Section 4.3). Signs in DGS may be one- or two-handed, and it is possible for each hand to articulate a different sign, so when these complex signs occur, we combine the two glosses into a single token, separated by “||”. For example, the token \$INDEX1\* || CAT1B\* indicates that the participant simultaneously signed \$INDEX1\* with their right hand and CAT1B\* with their left hand.

Table 2 shows the eight types of metadata included in MY DGS – ANNIS, six of which are available for each transcript, and three for each individual participant.

<sup>4</sup>Mouthings are provided in German for both versions, as they relate directly to articulation of German words and are therefore not suited for translation.

<sup>5</sup><https://www.sign-lang.uni-hamburg.de/hamnosys/input/>

<sup>6</sup>Further details of the Public DGS Corpus annotation conventions can be found in Konrad et al. (2022).

### 2.2. Annis Query Language (AQL)

Using AQL it is possible to query just one annotation tier or to make arbitrarily complex queries which refer to multiple annotation and metadata tiers. MY DGS – ANNIS provides a number of simple examples which users can use as a basis for creating their own queries. However, these cannot cover all possible combinations, and users have not always found it easy to work from these examples to create the queries which they need for their research. The main query types used in MY DGS – ANNIS are:

- regular expression search of the text associated with an element
- links between items in different tiers
- collocation distances between items in the same tier
- metadata

Each item in an AQL query must be linked to at least one other item. To facilitate this, each query item is automatically assigned a sequential number which can be used to refer to it later in the query. For example in Query 1, Gloss and English are connected using identifiers automatically assigned to them: #1 refers to the Gloss item and #2 to English. The identifiers can also be explicitly assigned, and we describe this process in section Section 4.1.

- (1) Gloss=/CAT.\* / & English=/.\* [Cc]at .\* / & #1 ->ident #2

Collocation distances are expressed using the dot (.) or caret (^) operators, followed by the tier name, then optionally by two numbers which specify the minimum and maximum distances. An example can be seen in Query 8.

Some examples of AQL searches in MY DGS – ANNIS can be found in Isard and Konrad (2022), and the full AQL manual is available online<sup>7</sup>. In Section 4 we show how the Query Wizard allows users to build up complex queries from smaller building blocks without the need to know the details of AQL syntax.

### 2.3. ANNIS Interface

Queries created in the Query Wizard can be opened directly in MY DGS – ANNIS (see Section 3). Figure 1 shows the MY DGS – ANNIS interface with the AQL query input window on the top left and the query results on the right for Gloss tokens with the gloss name cat (see Section 4 for a discussion of this query). Each result contains

<sup>7</sup><https://korpling.github.io/ANNIS/4.11/user-guide/aql>

Annotation	Description
Gloss	subtypes or types used to lemmatize tokens
GlossType	parent types
HamNoSys	HamNoSys notations of type citation forms
Mouth	mouthings or mouth gestures
Translation	for each utterance

Table 1: Annotation tiers in *MY DGS – ANNIS*

Metadata	Description	Refers to
TranscriptId	the unique identifier for the transcript	Whole Transcript
Region	where the transcript was recorded	
RegionCode	a shorter code for the region	
Date	date of the recording	
Theme	the task given to the participants for this transcript	
Keywords	a list of the topics discussed in this transcript	
Name	the unique (anonymous) identifier for each participant	Participant
AgeGroup	one of a set of four age categories	
Gender	the gender declared by each participant at the time of recording	

Table 2: Metadata included in *MY DGS – ANNIS*

three tabs which can be independently opened or closed; the first shows the five visible annotation tiers, with query results highlighted in red, the second the video for the transcript which can be played by clicking on any token or on the play button, and the third shows clickable links to another corpus portal, *MY DGS – annotated* (Konrad et al., 2024).

Figure 2 shows a view where we have zoomed in on the query result window, showing one match for the gloss *CAT1A\**, highlighted in red. *GlossType* and *HamNoSys* are also highlighted as they are directly linked to *Gloss* as alternative representations of the same gloss, while *Mouth* and *English* are independent tiers whose tokens can have different durations from the *Gloss* tokens. Unlike in Figure 1, the links tab has been opened in addition to the annotation and video tabs, providing links to the *MY DGS – annotated* Viewer and list of sign types.

### 3. The Query Wizard

The Query Wizard interface is a web application developed by us and written in JavaScript, that allows users to create a query by creating and linking smaller building blocks. It is available in English for creating queries for the English version of *MY DGS – ANNIS* and in German for the German version of the corpus. All examples in this article are shown for the English interface and corpus.

A user can create a block for any of the annota-

tion tiers, and the search can then be refined by the addition of search text. The options available for the text search depend on the tier selected, with the *Gloss* and *GlossType* tiers having the most additional options due to their more complex syntax as described in Section 2.1. Once a query has been generated, the user can click a button to open the query directly in *MY DGS – ANNIS*.

Figure 3 shows the initial state of the Query Wizard interface. The options available are to add a new block for an annotation tier or for a chosen metadata type. At this point, there is nothing displayed in the AQL query box at the top, and the button for opening the query in *MY DGS – ANNIS* is therefore greyed out. There is a button to create connections between elements and a display for the list of connections between annotation elements, but both are empty, as no elements have yet been created.

When the user has selected a tier and clicked the add button, a new block appears, where they can refine the search as shown in Figure 4. This can be done by adding text in the search box, and if desired, constraining the search with further options. If they want to find glosses with a particular gloss name, they can enter free text in the search box, and constrain whether the gloss name should exactly match the text entered, start with the text, or contain the text.

Each annotation or metadata block can be temporarily excluded from the query by unchecking the

The screenshot shows the ANNIS interface with the following components:

- Search Bar:** Query: G1#Gloss=(/\*)CAT[0-9][A-Z]?/ & #G1 ...
- Results List:**
  - 1 DGS-Corpus-r3-en > 1176846
    - Annotations: PersonA::English: Later in the evening, she really wanted to go home because of her cat.
    - PersonA::GlossType: WHISKERS1A\*
    - PersonB::Gloss: CAT1A\*
    - PersonB::HamNoSys: ...
    - PersonB::Mouth: katze
  - 2 DGS-Corpus-r3-en > 1176846
    - Annotations: PersonA::English: I saw... until eleven o'clock.
    - PersonA::GlossType: SORAL\*
    - PersonB::Gloss: SORAL\*
    - PersonB::Mouth: all jhr
    - PersonB::English: Then I drove her home.
    - PersonB::GlossType: TO-BRING1A\*
    - PersonB::Gloss: TO-BRING1A\*
    - PersonB::HamNoSys: ...
    - PersonB::Mouth: nach hause bringen
  - 3 DGS-Corpus-r3-en > 1176846
    - Annotations: PersonA::English: She was scared on New Year's Eve, so she hid.
    - PersonA::GlossType: MAGIC3\*
    - PersonB::Gloss: MAGIC3\*
    - PersonB::HamNoSys: ...
    - PersonB::Mouth: verstecken
  - 4 DGS-Corpus-r3-en > 1176846
    - Annotations: PersonA::English: That's nice.
    - PersonA::GlossType: BEST-DECLINE1\*
- Video Player:** Shows two people sitting in a studio with a blue background and a 'MEINE DGS' logo.

Figure 1: MY DGS – ANNIS showing the query results for tokens with gloss name CAT and the metadata “Theme”.

The zoomed-in view shows the following details for the token CAT1A\*:

- Annotations:**
  - PersonB::English: The cat got really excited, “Grandma is back again.”
  - PersonB::GlossType: WHISKERS1A\*
  - PersonB::Gloss: CAT1A\*
  - PersonB::HamNoSys: ...
  - PersonB::Mouth: katze
- Video Player:** Shows the same studio setting as Figure 1.
- Links to MY DGS — annotated**
- 1176846**
- Person MY DGS — annotated Viewer Type right Hand Type left Hand**

Person B	CAT1A*	CAT1A*
Person B	HAPPY1	HAPPY1
Person B	GRANDMA3*	GRANDMA3*
Person B	NOT2	NOT2
Person B	ONCE-MORE1A	ONCE-MORE1A
Person B	PRESENT-OR-HERE1	PRESENT-OR-HERE1

- DGS-Corpus-r3-en > 1176846**
- Annotations:**
  - PersonA::English: She was scared on New Year's Eve, so she hid.
  - PersonA::GlossType: MAGIC3\*, SINDE1A\*, WHISKERS1A\* || SINDE1A\*, FEAR1A\*

Figure 2: Zoomed-in view of Figure 1 showing one specific result with all tabs expanded.

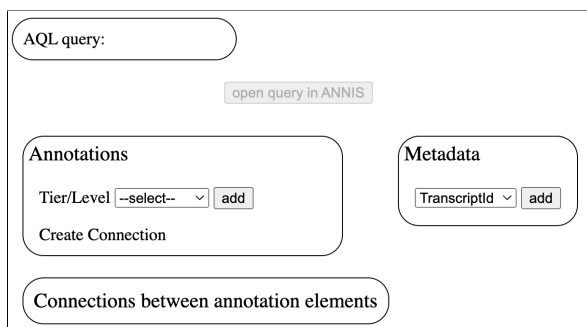


Figure 3: The start interface for the Query Wizard

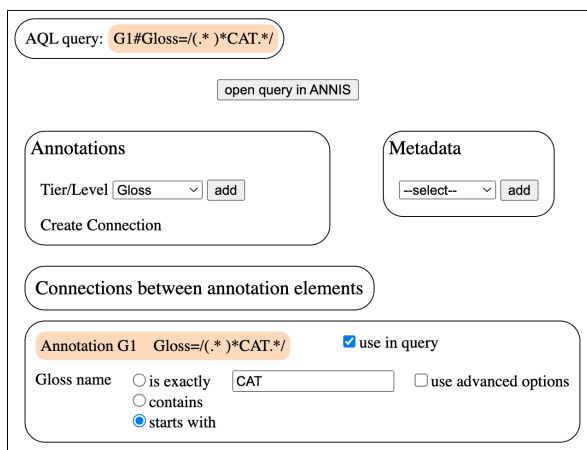


Figure 4: Interface with a Gloss block where the gloss name starts with the string CAT

“use in query” checkbox; if the box is later checked again, all search parameters previously entered are still active. As edits are carried out in a block, the AQL query display at the top of the interface changes accordingly.

We mentioned in Section 2.2 that each item in an AQL query is automatically assigned a sequential number which can be used to refer to it later in the query. It is also possible to explicitly assign an identifier to each item in a query, and the Query Wizard does this, giving each item a code which starts with a letter representing the annotation tier (G for Gloss, GT for GlossType and so on) and a number which is incremented every time an item from the same tier is created. In Figure 4 there is a single Gloss item so it receives the identifier G1. This is used to refer to it in the AQL query, but also every time the item is referenced elsewhere in the interface.

Figure 5 shows the search block from Figure 4 with the “use advanced options” checkbox selected. When this checkbox is first selected, the checkboxes for “all lexical variants” and “all phonological variants” are selected, and the “allow supertypes” and “allow modified” checkboxes are set to “all”. To avoid the creation of invalid searches, the lexical variants box only becomes active when the user

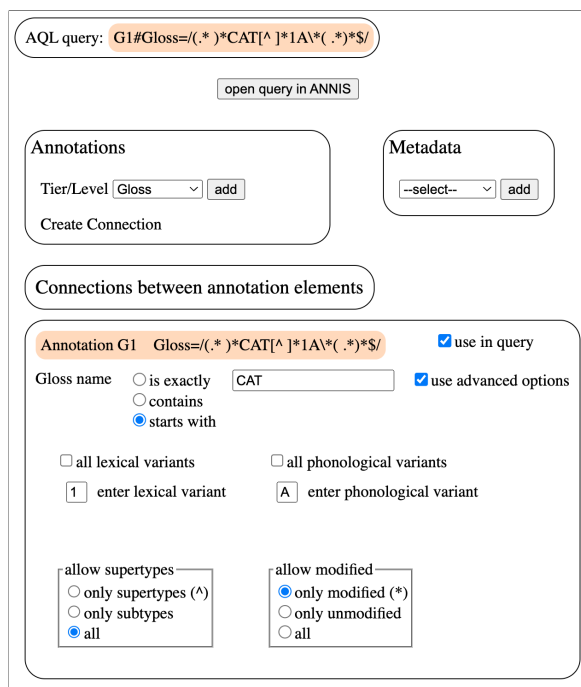


Figure 5: Interface with a Gloss block where the gloss name starts with the string CAT, with advanced options selected

has entered some text into the search box, and the phonological variants box only becomes active when a lexical variant has been entered. In Figure 5 the search is restricted to lexical variant 1, phonological variant A and only tokens which diverge from the citation form.

When a second annotation block is created, it is assigned a different colour in order to help the user to identify which part of the AQL query comes from which block. The main AQL query display temporarily becomes blank, because all annotation items in a valid AQL query must be linked to one another in some way. If there are annotation items which are not yet connected, a dropdown list of the items becomes available, using the identifiers which have been assigned by the Query Wizard. There are two kinds of connections: links between tokens which occur at the *same time* on different tiers, and collocation distances between tokens on the same tier. Collocation searches can find tokens *before* or *after* a token on the same tier, or permit both directions with the option *near*. In addition, it is possible to constrain the collocation distance with minimum and maximum values.

Figure 6 shows how the creation of a Gloss block and an English block has made available a dropdown menu to connect the two. Once a connection has been configured, it can be added and then appears in the list of connections and is integrated into the main AQL query, as can be seen in Figure 7.



Figure 6: Interface for connecting two blocks.

Figure 7: Interface with a Gloss G1, English G2 and a *same time* connection between them.

## 4. Examples

### 4.1. AQL Regular Expressions and Gloss Syntax

A simple query might be, for example, to search the corpus for all translations which contain the word “cat”, which can be expressed in AQL as shown in [Query 2](#) and for which we find 194 matches.

(2) `English=/. * [Cc]at .*/`

In this case, the user needs basic knowledge of the Public DGS Corpus annotations, plus simple AQL and regular expression syntax:

- English translations are in a tier named “English”

- AQL regular expression search is denoted by a query between two forward slashes (“/”)
- In a regular expression, “. \*” matches 0 or more characters of any kind
- In a regular expression, “[Cc]” matches either “C” or “c”

Now, if we want to instead search for tokens with the gloss word `CAT`, we could try the query in [Query 3](#). As before, this requires some knowledge of the DGS Corpus annotations, AQL, and regular expression syntax.

(3) `Gloss=/CAT.* /`

[Query 3](#) gives us 119 matches, which seems plausible given the 194 matches for [Query 2](#), but if we examine them, we discover that only 75 are actually matches for varieties of `CAT`, while 25 are varieties of `CATHOLIC`, 13 `CATHEDRAL`, 4 `CATASTROPHE` and 2 `CATTLE`.

As described in [Section 2.1](#), lexical variants of a sign are indicated with different digits after the gloss word. A next attempt would therefore be [Query 4](#), which does indeed give 75 results – success!

(4) `Gloss=/CAT[0-9].* /`

However, we then remember that signs performed with the left hand are prefixed with “||” (as explained in [2.1](#)), and with [Query 5](#) we indeed discover 4 instances of `CAT1A*` performed with the left hand, and one of `CAT1B*` performed with the left hand co-articulated with `$INDEX1*` performed with the right hand.

(5) `Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.* /`

By this point, the regular expression has already become fairly complicated, and if we wanted to further restrict this query to only supertypes or subtypes, or to only modified or unmodified glosses (see [Section 2.1](#) for explanations), it would become significantly more so.

In the Query Wizard we only need to write the gloss name `CAT`, and select the button “is exactly”, and the regular expression is created automatically, as shown in [Figures 6](#) and [7](#).

### 4.2. Corpus Metadata

After finding all the tokens with gloss name `CAT`, a user may be curious in which corpus themes and in which transcripts these tokens most frequently occurred. In order to find this out, they need to add two metadata items to the query. To build this query manually, the user would have to know not only the type structures described above, but also the exact names of the two metadata types and the syntax for linking them. Each metadata item must be linked from an annotation item using the string “@\*”, as shown in [Query 6](#).

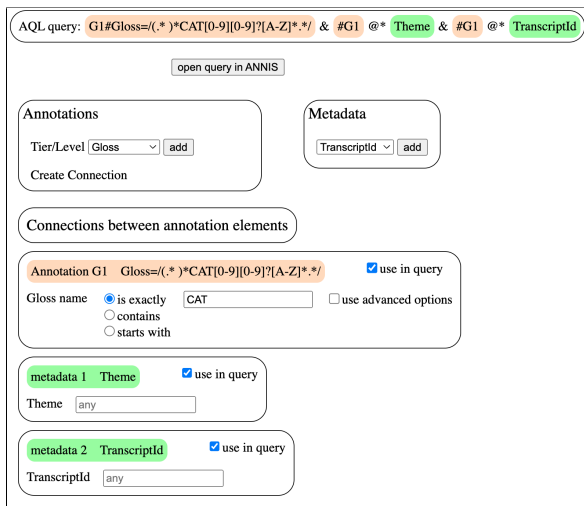


Figure 8: Interface with a Gloss block with gloss name CAT plus metadata Theme and TranscriptId

(6) `G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*./ & #G1 @* Theme & #G1 @* TranscriptId`

In Query Wizard this query can be created simply by creating the annotation block for the Gloss CAT as shown before, and two metadata blocks, one for Theme and one for TranscriptId, as in Figure 8. The results of opening this query in MY DGS – ANNIS are shown in Figures 1 and 2. The frequency analysis tab of MY DGS – ANNIS, shown in Figure 9, can then be used to discover that the CAT glosses occur most frequently in the “Sylvester and Tweety” task, where participants are asked to retell the popular cartoon story, but also often in discussions on specific “Subject Areas” and occasionally in 5 other themes, including “Experience of Deaf Individuals”.

Alternatively, a user might want to search only for results from participants from the age group “18–30”. This is more complicated because of the way that the metadata is stored internally in the ANNIS database. In order to search metadata specific to one participant it is necessary to create a query which explicitly links each person’s tokens to their metadata, as shown in Query 7. Again this query can be simply created in the Query Wizard by creating a gloss block for CAT and a metadata block for age group, as shown in Figure 10.

(7) `(G1#PersonA:Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*./ @* PersonA:AgeGroup="18-30") | (G1#PersonB:Gloss=/(.*)*CAT.*? @* PersonB:AgeGroup="18-30")`

### 4.3. Collocation Distances

In the final example, shown in Figure 11, the user has created two Gloss blocks. The first, G1, searches as before for all tokens with gloss name CAT. The second, G2, does not specify any search text, and has a collocation distance from G1 of 1 to

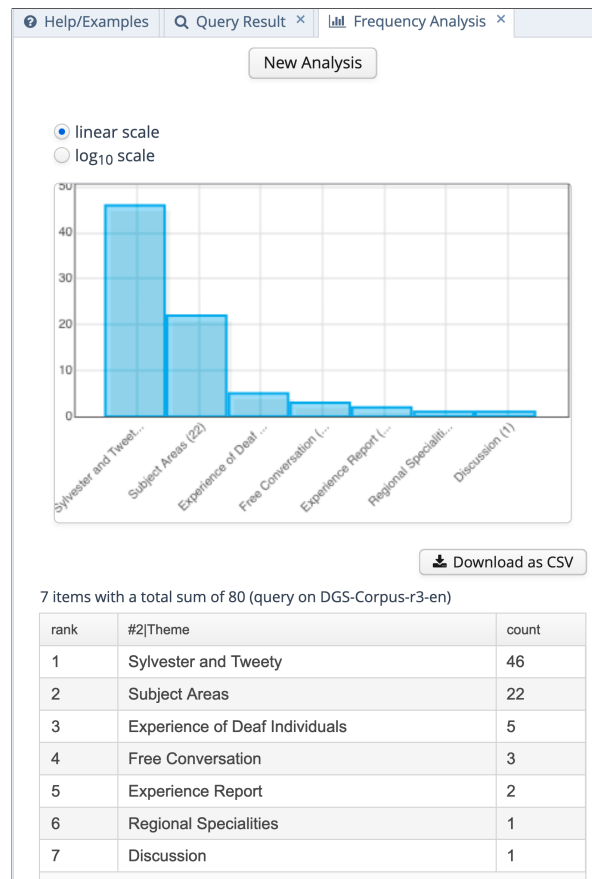


Figure 9: MY DGS – ANNIS frequency analysis showing in which themes the tokens with gloss name CAT appear most frequently

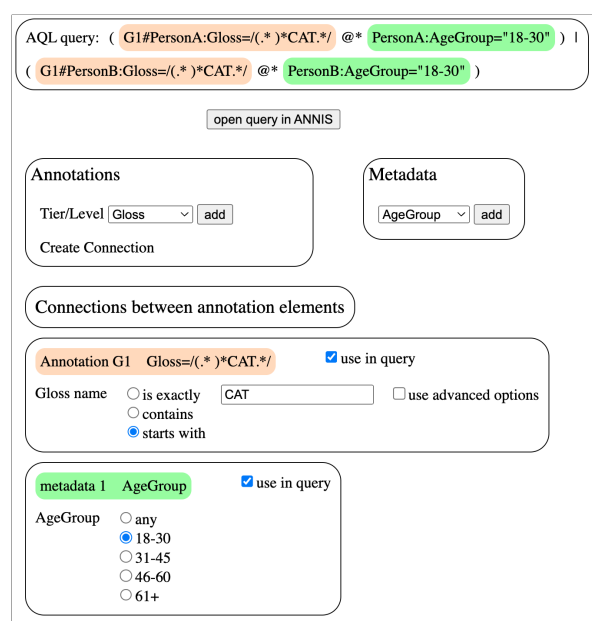


Figure 10: Interface with a Gloss block with gloss name CAT plus metadata AgeGroup=18-30

AQL query: `G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & G2#Gloss & #G1 ^Gloss,1,2 #G2`

[open query in ANNIS](#)

**Annotations**

Tier/Level --select-- add

Create Connection

**Metadata**

--select-- add

**Connections between annotation elements**

G1 from 1 to 2 near G2 delete

**Annotation G1** Gloss=/(.\*)\*CAT[0-9][0-9]?[A-Z]\*.\*/  use in query

Gloss name  is exactly   use advanced options

contains

starts with

**Annotation G2** Gloss  use in query

Gloss name  is exactly   use advanced options

contains

starts with

Figure 11: Interface with two Gloss blocks with a collocation distance of 1 to 2.

2 in either direction, and the AQL query is shown in Query 8.

(8) `G1#Gloss=/(.*)*CAT[0-9][0-9]?[A-Z]*.*/ & G2#Gloss & #G1 ^Gloss,1,2 #G2`

When this query is opened in the frequency analysis tab of ANNIS (see Figure 12), we can see that the most frequent collocations are special signs, including productive signs and pointing gestures. The most frequent lexical sign is GOOD1, which leads us to a tentative and humorous conclusion that the corpus participants are well-disposed towards cats.

## 5. Conclusions and Future Work

We have introduced the new Query Wizard for *MY DGS – ANNIS* and shown how it simplifies the process of building queries in the ANNIS AQL query language for the DGS Corpus. It allows users to build queries out of small building blocks, and helps them to understand how the queries are built up. It removes the burden of regular expression building from users, and means that they do not have to remember the spellings of annotation tier and metadata names. It also allows users to select from the valid sets of metadata options. While user studies are still ongoing, initial feedback has been very favourable.

New corpus releases will be published as separate datasets in *MY DGS – ANNIS*. These new datasets may introduce new tiers or change structural aspects to account for new corpus (meta)data

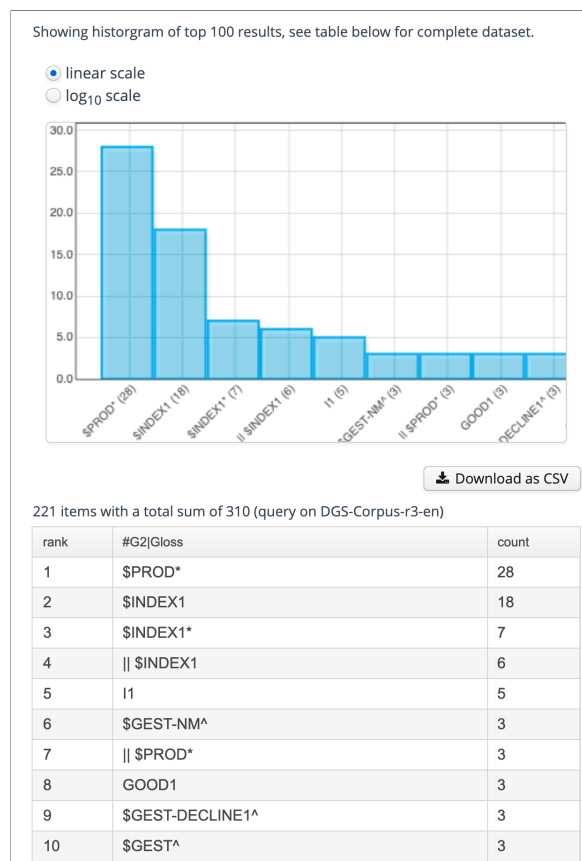


Figure 12: *MY DGS – ANNIS* frequency analysis of signs with a collocation distance of 1 to 2 from signs with gloss word CAT

and improvements to the ANNIS software. The Query Wizard will allow users to choose the desired corpus release and will adjust its query outputs accordingly.

There are also a number of features yet to be added to the Query Wizard. These include negated searches and fine-grained control over handedness of sign execution. Entering HamNoSys may be further improved by integrating the HamNoSys Builder interface more directly into the Query Wizard or supporting the use of HamNoSys character names, such as *hampinch12open*. Options could also be included to allow users to search special classes of signs such as numbers and fingerspellings.

## 6. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

## 7. Bibliographical References

- Stefanie Dipper. 2015. [Annotierte Korpora für die Historische Syntaxforschung: Anwendungsbeispiele anhand des Referenzkorpus Mittelhochdeutsch](#). *Zeitschrift für germanistische Linguistik*, 43(3):516–563.
- Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. [Extending the Public DGS Corpus in size and depth](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).
- Amy Isard and Reiner Konrad. 2022. [MY DGS – ANNIS: ANNIS and the Public DGS Corpus](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).
- Reiner Konrad, Thomas Hanke, Amy Isard, Marc Schulder, Lutz König, Julian Bleicken, and Oliver Böse. 2024. [Corpus à la carte – improving access to the Public DGS Corpus](#). In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy. European Language Resources Association (ELRA).
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. [Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions](#). Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Thomas Krause and Amir Zeldes. 2016. [ANNIS3: A new architecture for generic corpus query and visualization](#). *Digital Scholarship in the Humanities*, 31(1):118–139.



# Investigating Motion History Images and Convolutional Neural Networks for Isolated Irish Sign Language Fingerspelling Recognition

Hafiz Muhammad Sarmad Khan<sup>\*†</sup> , Irene Murtagh<sup>\*†</sup> , Simon D McLoughlin<sup>†</sup> 

<sup>\*</sup>ADAPT: Centre for AI-Driven Digital Content Technology, <sup>†</sup>Technological University Dublin  
Dublin, Ireland

sarmad.khan@adaptcentre.ie, irene.murtagh@adaptcentre.ie, simon.d.mcloughlin@tudublin.ie

## Abstract

The limited global competency in sign language makes the objective of improving communication for the deaf and hard-of-hearing community through computational processing both vital and necessary. In an effort to address this problem, our research leverages the Irish Sign Language hand shape (ISL-HS) dataset and state-of-the-art deep learning architectures to recognize the Irish Sign Language alphabet. We streamline the feature extraction methodology and pave the way for the efficient use of Convolutional Neural Networks (CNNs) by using Motion History Images (MHIs) for monitoring the sign language motions. The effectiveness of numerous powerful CNN architectures in deciphering the intricate patterns of motion captured in MHIs is investigated in this research. The process includes generating MHIs from the ISL dataset and then using these images to train several CNN neural network models and evaluate their ability to recognize the Irish Sign Language alphabet. The results demonstrate the possibility of investigating MHIs with advanced CNNs to enhance sign language recognition, with a noteworthy accuracy percentage. By contributing to the development of language processing tools and technologies for Irish Sign Language, this research has the potential to address the lack of technological communicative accessibility and inclusion for the deaf and hard-of-hearing community in Ireland.

**Keywords:** Motion History Images, Irish Sign Language Recognition, Convolutional Neural Networks

## 1. Introduction

Sign Languages (SLs), expressed visually through gestures within a three-dimensional signing space, and without a written form serve as the principal mode of communication for numerous deaf and hard-of-hearing communities in their daily interactions. The fact that sign languages are often overlooked by current natural language processing and machine translation technologies exacerbates the existing communication challenges faced by the estimated 72 million deaf individuals worldwide (Murtagh et al., 2022; Murtagh, 2021). Irish Sign Language (ISL) maintains a unique place in the sign language landscape, serving as the principal means of communication for Ireland's deaf and hard-of-hearing community (Leeson and Saeed, 2012). Irish Sign Language (ISL) constitutes a gestural mode of communication devoid of written or spoken articulation. It serves as the primary means of interaction for approximately 5,000 Deaf individuals within Ireland. An additional 40,000 hearing individuals engage with ISL, exhibiting a spectrum of usage frequency from regular to occasional within the country (School of Linguistic, Speech and Communication Sciences, 2016) (Irish Deaf Society). Notwithstanding its cultural and linguistic significance, ISL like numerous other sign languages across the world faced with technologi-

cal constraints due to its visual and spatial properties. Human-Computer Interaction (HCI) is strongly linked to advancements in computer vision, with recognition being a focus for research. The failure to integrate sign languages into modern technologies has hindered the development of accessible information and services for the ISL community, compounded by the challenge of the limited availability of comprehensive datasets for training and evaluating AI models in computational processing. The purpose of this research is to utilize an Irish Sign Language dataset and explore the effectiveness of sophisticated neural network frameworks in recognizing ISL hand motions from motion history images. The efforts are ongoing in the development of a computational system that will automatically annotate sign language data, hence improving communication accessibility and inclusivity for the ISL community.

The paper's outline is structured as follows: Section 2 offers an overview of relevant research in sign language recognition. In Section 3, the proposed methodology is presented, covering the dataset description, data augmentation techniques, and experimental architectures. Section 4 elaborates on the experimental results, and Section 5 culminates in the conclusion.

## 2. Related Work

In recent years, the evolution of artificial intelligence (AI) and computer vision has fueled dramatic advances in sign language understanding, solving major issues for those with deaf and hard-of-hearing communities (LeCun et al., 2015). This section presents the progress of sign language recognition, with an emphasis on the critical role of deep learning approaches in improving accessibility and communication among the deaf and hard-of-hearing populations.

One of the foundational contributions to this field was proposed by Mathieu De Coster et al. (De Coster et al., 2021), by presenting a novel approach to enhance the performance of the Video Transformer Network (VTN) for isolated sign recognition leveraging multi-modal inputs, including human pose key points and hand crops, extracted from RGB videos. Their adaptation addresses the challenge of limited labeled data available for sign language recognition by enriching the model's input with pre-processed information that captures essential features of sign language, such as hand shapes and body movements. The methodology demonstrated a significant improvement in sign recognition accuracy, achieving 92.92% on the AUTSL dataset, underscoring the potential of combining pose estimation and self-attention mechanisms in deep learning models for more accurate and interpretable sign language recognition. This research was conducted under the SignON project (Sig), funded by the European Union's Horizon 2020. Mathias Müller et al. (Müller et al., 2022), present the inaugural shared task for automatic translation between signed and spoken languages, specifically focusing on Swiss German Sign Language (DSGS) to German and vice versa. This pioneering effort marks a significant departure from the traditional text-to-text machine translation, necessitating the processing of visual information such as video frames or human pose estimation. The task attracted seven teams, all participating in the DSGS-to-German track, showcasing state-of-the-art techniques. Additionally, it generated the first publicly available dataset of system outputs paired with human evaluation scores for sign language translation, thereby setting a foundational benchmark for future research in this emergent field. Neha Deshpande et al. (Deshpande et al., 2022) investigate the use of convolutional neural networks (CNNs) for facial expression recognition in sign language videos, targeting Ekman's six basic expressions (fear, disgust, surprise, sadness, happiness, anger) plus a neutral category. They enhance the performance of pre-trained general facial expression models through fine-tuning, data

augmentation, class balancing, and image pre-processing. Their method, validated using K-fold cross-validation, significantly improves accuracy on sign language datasets, showcasing the effectiveness of CNNs in sign language facial expression recognition and contributing valuable insights to the field. In the pioneering work, Wong et al. (Wong et al., 2022) introduce a novel Hierarchical Sign I3D model (HS-I3D), significantly advancing the field of sign spotting in continuous sign language videos. By innovatively applying a hierarchical spatiotemporal network architecture to learn coarse-to-fine sign features, their approach adeptly captures signs at varying temporal levels, leading to more accurate sign localization. Evaluated on the ChaLearn 2022 Sign Spotting Challenge - MSSL track, the HS-I3D model notably achieved a state-of-the-art 0.607 F1 score, marking it as the competition's top-performing solution. This achievement not only demonstrates the model's effectiveness in identifying and localizing signs with high precision but also emphasizes the utility of incorporating random sampling techniques during model training. (Hsieh et al., 2010) introduced an adaptive approach for hand gesture recognition in human-machine interactions. The novel approach, which integrated an adaptive skin color algorithm with facial recognition algorithms, demonstrated outstanding accuracy even in low-light circumstances and complicated backdrops. This research conducted experiments in which five persons made 250 hand motions at different distances from the webcam. The proposed system demonstrated its practicality and usefulness in real-world applications, with an average accuracy of 94.1% and a processing time of 3.81 milliseconds per frame. This study lays the framework for future advances in sign language recognition algorithms. (Yalçinkaya et al., 2016) highlighted the importance of sign language recognition in improving communication for those with speech and hearing impairments. Their system, which used Motion History Images (MHI) and a nearest neighbor approach, obtained an excellent classification accuracy of 95%, demonstrating the capacity of machine learning to bridge communication gaps. This demonstrates AI's revolutionary influence on increasing accessibility for underserved populations. The implementation of convolutional neural networks (CNNs) has accelerated developments in the recognition of sign language. (Barbhuiya et al., 2021) used CNN frameworks to extract and categorize characteristics in sign language motions, resulting in excellent classification accuracy. Using pre-trained CNN models like "AlexNet" and "VGG-16," they demonstrated the usefulness of deep neural networks in practical applications of sign language recognition systems. Quantitative evaluations demonstrate the efficacy of the

CNN-based method, with the model achieving high accuracy rates in sign language categorization of 99% when using random validation and 70% when utilizing leave-one-out validation. Simultaneously, (Wadhawan and Kumar, 2020) made significant advances in deep learning-based CNNs for sign language identification by representing static signs. Through testing and analysis, they were able to get exceptional training precision, outperforming earlier methods and creating new opportunities for identifying a wider variety of hand signals. The suggested approach achieved remarkable training accuracy of 99.90% and 99.72%, respectively. This demonstrates how AI-driven methods for sign language processing are evolving and improving. Bantupalli et al. (Bantupalli and Xie, 2018) proposed an innovative technique to address communication challenges encountered by individuals who have speech impairments. Their research focuses on the creation of a vision-based application for sign language translation into text. Using current advances in deep learning and computer vision, they extracted important temporal and spatial information from video sequences. They specifically used Inception to recognize spatial features and a Recurrent Neural Network to analyze temporal data. The experiment yielded good results, with an average accuracy of 90% with the softmax layer and 55% with the pooling layer. The study emphasizes the transformative potential of technology-driven solutions in overcoming societal difficulties, as well as the significance of interdisciplinary collaboration in fostering social innovation. R.S. Sabeenian et al. (Sabeenian et al., 2020) investigated the challenges linked to speech impairment affecting communication via speech and hearing. Despite the growing usage of sign language as an alternate communication tool, non-signers continue to face a hurdle in communicating effectively with signers. Making use of recent advances in computer vision and deep learning, the authors concentrated on creating a deep learning-based application for translating sign language into text. Their method used a proprietary Convolutional Neural Network (CNN) to recognize signs in video frames, with the MNIST dataset used for model training. The constructed model attained 93% accuracy, indicating its usefulness in sign language identification and translation. Dongxu Li et al. (Li et al., 2020) developed the Word-Level American Sign Language (WLASL) (Li, 2020) video collection, which includes over 2000 words performed by 100+ signers, to overcome the limitations of existing sign language datasets. Their research enabled testing with deep learning methods for word-level sign identification, contrasting holistic visual appearance-based and 2D human pose-based approaches. Furthermore, they suggested a novel Pose-based Temporal Graph

Convolution Networks (Pose-TGCN) method to improve pose-based recognition. Both approaches produced equivalent results, with up to 62.63% top-10 accuracy on 2000 words/glosses, demonstrating the dataset's importance in improving sign language recognition research.

In spite of these developments, there is still a significant shortfall of research using this approach for Irish Sign Language (ISL). By filling up this gap, future research will have the chance to improve accessibility and inclusion for members of the ISL community.

### 3. Proposed Methodology

This section covers the architectures used for sign language recognition, as well as the dataset utilized and data augmentation aspects.

#### 3.1. Dataset

In this research, we utilized the Irish Sign Language hand-shape dataset (ISL-HS) (Oliveira et al., 2017), which consists of real hand images. The ISL-HS dataset consists of 23 static gestures representing English alphabet signs and three dynamic motions (J, X, and Z). The recording method was led by the dataset documentation, with six people (three men and three women) practicing fingerspelling of the ISL alphabet, with each action recorded three times. The videos were captured at 30 frames per second (fps) and 640x480 pixels, for a total of 468 recordings. From these films, 52,688 frames were retrieved for static forms and 5,426 frames for dynamic motions, for a total of 58,114. We used both static and dynamic form images in this investigation.

#### 3.2. Preprocessing and Augmentation Strategies

##### 3.2.1. Data PreProcessing

In our approach to the Irish Sign Language hand shape dataset, the first stage entailed converting the dataset into motion history images (MHIs). To accomplish this, we employed a tailored Python script that made use of computer vision libraries like OpenCV. This script systematically traversed through the sequence of images, computing the absolute variance between successive frames. Subsequently, it updates the Motion History Image by considering a predefined motion threshold and persistence parameter. The resultant MHI effectively encapsulates the temporal motion information, ultimately yielding the final MHI image. Following the generation of MHI, the dataset now comprises a total of 18 images for each letter of the alphabet, culminating in a robust set of 468 images. Figure 1

showcases the original dataset image and its corresponding MHI representation. The equation for calculating the Motion History Image (MHI) is given by:

$$MHI(x, y) = \begin{cases} \tau & \text{if } |I_t(x, y) - I_{t-1}(x, y)| > mv\_thresh \\ MHI(x, y) - 1 & \text{otherwise} \end{cases} \quad (1)$$

where  $MHI(x, y)$  represents the pixel value at location  $(x, y)$  in the MHI,  $I_t(x, y)$  is the pixel value at location  $(x, y)$  in the current frame  $t$ ,  $I_{t-1}(x, y)$  is the pixel value at location  $(x, y)$  in the previous frame  $t - 1$ ,  $\tau$  is the persistence parameter, and  $mv\_thresh$  is the motion threshold parameter.

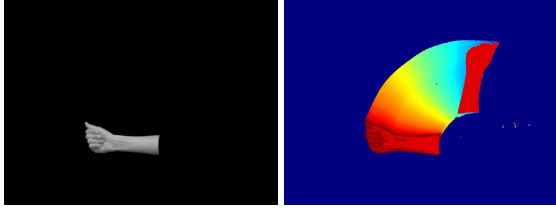


Figure 1: Left: Original image from the dataset. Right: Motion History Image (MHI) representation.

### 3.2.2. Data Augmentation

After developing the motion history images (MHIs), we employed data augmentation to expand the dataset and improve the robustness of frameworks. Due to the scarcity of the dataset, we chose data augmentation via synthesis (Keskin, 2023; Jo et al., 2017). This process involved developing new iterations of the images using various augmentations. The goal was to incorporate unpredictability into the dataset, which would improve the model’s ability to generalize to new data and increase its effectiveness in real-world applications. Specifically, augmentations such as additive Gaussian noise and multiplication were used to bring variation into the images. Additionally, affine transformations such as scaling and zooming were used to imitate changes in viewpoint and orientation. By applying various augmentation techniques to the 468 original images, we generated new iterations of these images, significantly expanding our dataset to a total of 7020 images. This increase has resulted in a substantial enhancement, providing 270 images for each class, thereby reinforcing the robustness of our frameworks. Of these images, 80% were designated for training and validation while the remaining 20% were reserved for testing in order to increase the resilience of the framework. These augmentations contributed to a more diversified and complete dataset, allowing for more efficient training of models and improved sign language recognition performance. Overall, data augmentation was critical in improving the quality and variety of our dataset,

Description	Total Images	Images per Class
Original Images	58112	2235
Post-MHI Images	468	18
Post-Augmentation Images	7020	270
Training Set	4420	170
Validation Set	1196	46
Test Set	1404	54

Table 1: Dataset Overview: Table presents key metrics at various processing stages, including total images and images per class.

resulting in enhanced model performance and generalization capabilities for detecting Irish Sign Language motions. Figure 2 illustrates examples of augmented images representing each letter in MHI format and the dataset composition, detailing total images, images per class, and the train-test split at different processing stages, is summarized in Table 1.

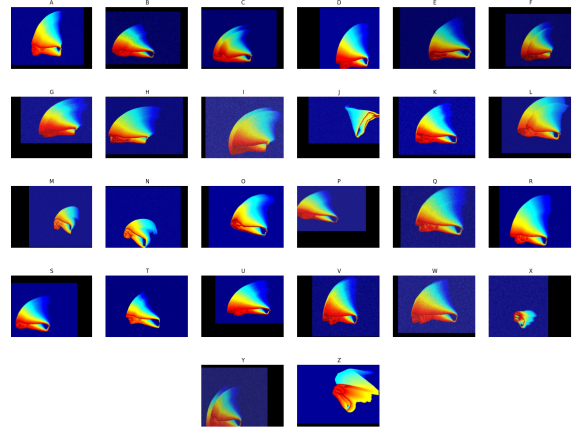


Figure 2: Sample Augmented Image for each Letter

### 3.3. Proposed Architectures

A multitude of deep learning algorithms exist, offering various capabilities and applications. However, among these algorithms, Convolutional Neural Networks stand out prominently in the field of computer vision. Researchers often choose CNNs for image classification tasks due to their ability to effectively analyze images as input and output probability values or class labels (Putzu et al., 2020). This capability makes CNN architectures highly suitable for addressing challenges in picture categorization, such as identifying objects, patterns, or gestures within images. In our research, we leverage the power of CNN architectures for sign language recognition



using motion history images. As illustrated in Figure 3, the MHI sign language recognition pipeline encompasses a sequence of stages from collective dataset acquisition, through preprocessing and data augmentation, to feature extraction state-of-the-art CNN networks, and finally to recognition and probability testing yielding the results. As shown in Table 2, we explore a variety of CNN architectures tailored to the unique properties of motion history images, aiming to enhance the accuracy and robustness of the Irish Sign Language recognition system.

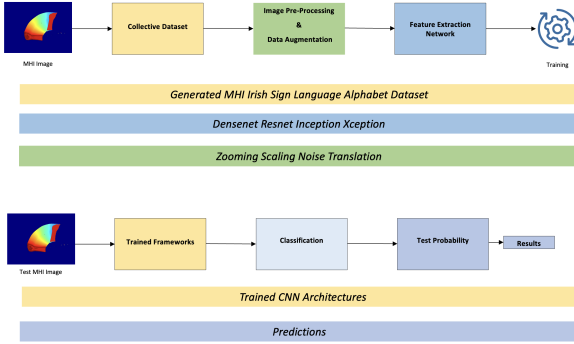


Figure 3: Flowchart depicting the training and testing pipeline for an Irish Sign Language recognition system using CNN architectures, with the top path illustrating the training phase on augmented MHI data, and the bottom path showing the testing phase leading to predictions.

Model	Architecture	Parameters
Resnet	50-V2	25.6 M
	101-V2	44.7 M
	152-V2	60.4 M
Xception	Xception	22.9 M
Densenet	121	8.1 M
	169	20.2 M
	201	14.3 M
Inception	V3	23.9 M
	Resnet-V2	55.9 M

Table 2: Frameworks utilized and Parameters

## 4. Experimental Results

### 4.1. Model Hyperparameters

TensorFlow and Keras libraries are employed to implement deep learning architectures. Additionally, image augmentation techniques are utilized, with the training epoch set to 20. Fine-tuning, a crucial step in model optimization, is also performed to adapt the pre-trained models. This process involves adjusting the parameters of the pre-trained

models to better suit the characteristics of the target dataset, thereby enhancing performance. A learning rate of 0.001 is utilized, along with a batch size of 16, and optimization is achieved using the ADAM optimizer with the "categorical cross-entropy" loss function. The softmax activation function is applied to the models. To expedite processing, the resolution is standardized to 160x120. Python scripts are executed on Google Colab, leveraging the Tesla K80 GPU for enhanced computational efficiency.

### 4.2. Evaluation Metrics

Accuracy, precision, recall, and F1-Score were used in the appraisal of the framework's performance.

Recall is the measure of how the model and algorithm predict True positives:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

where TP stands for True Positive and FN stands for False Negative.

Precision is determined by the ratio of properly identified true negative samples to the total number of outcomes, which includes both true negative and false positive results:

$$\text{Precision} = \frac{TN}{TN + FP} \quad (3)$$

where TN stands for True Negative and FP stands for False Positive.

The accuracy of the model is determined by the proportion of its predictions that are confirmed by testing:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TP, TN, FP, and FN have the same meanings as above.

F1-Score is a technique that is used to combine the accuracy and recall of the model and is also the harmonic combination of the model's recall and precision:

$$\text{F1 Score} = \frac{2TP}{2TP + FP + FN} \quad (5)$$

where TP, FP, and FN have the same meanings as above.

### 4.3. Results

The findings of our research show that different deep learning models for motion history image (MHI)-based sign language recognition have differing accuracy levels. Table 3 shows that

Densenet 121 attains the greatest accuracy of 90.38%, Densenet 201 closely behind with 90.10%, and Densenet 169 with 89.60%. Of all the examined frameworks, Densenet continually showed the optimal accuracy rates. The ability of Densenet frameworks to reliably identify Irish Sign Language alphabets from MHIs is demonstrated by this. Although Densenet architectures showed remarkable performance, other models also yielded encouraging outcomes. With ResNet variations (ResNet 101 V2, ResNet 50 V2, and ResNet 152 V2) varied from 82.24% to 85.62%, Xception attained an accuracy of 80.56%. Furthermore, 77.64% and 75.76% accuracy were attained using Inception ResNet V2 and Inception V3, respectively. These results imply that different deep learning architectures could potentially be applied successfully to sign language challenges. Notably, we maintained the same amount of hyperparameters for each model, including epoch, batch size, loss function, and optimizer. However, the framework’s particular needs and limitations should be taken into account while selecting a model architecture. Densenet frameworks, for example, have the maximum accuracy, but alternative models could offer a fair trade-off between accuracy and computing economy, which would make them more appropriate for some deployment scenarios. As demonstrated in Figure 4, the confusion matrix showcases the performance of the Densenet 201 framework across all classes, with accuracy observed along the diagonal where predicted values coincide with actual values.

Model	Accuracy
Densenet 121	90.38
Densenet 169	89.60
Densenet 201	90.10
Xception	80.56
Resnet 101 V2	85.62
Resnet 50 V2	82.84
Resnet 152 V2	82.24
Inception Resnet V2	77.64
Inception V3	75.76

Table 3: Evaluation Metrics

## 5. Conclusion

In conclusion, our research offers valuable insights into the application of deep learning methodologies for sign language recognition, particularly leveraging motion history images (MHIs). Despite the challenges posed by limited datasets, our research highlights the effectiveness of data augmentation techniques in enhancing model performance. By evaluating various state-of-the-art architectures on an

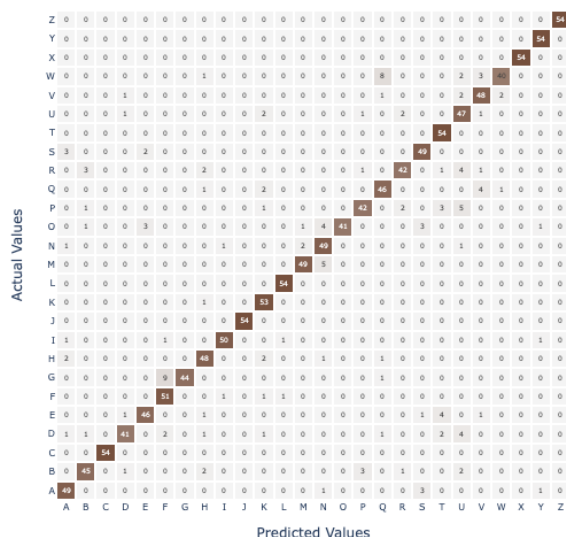


Figure 4: Confusion matrix illustrating the performance metrics on the ISL-HS test dataset

Irish Sign Language motion history images dataset, we have identified promising avenues for improving accessibility and inclusivity for the deaf and hard-of-hearing community. Looking ahead, we envision delving into more intricate challenges, such as sign language annotation and the development of automated annotation pipelines. Through further investigation into the efficiency of deep learning frameworks in computer vision, we aim to narrow the technological gap in sign language recognition, thus contributing to advancements in accessible communication technologies.

## 6. Acknowledgments

We would like to sincerely thank all the organizations who helped make this research project a reality. We are grateful for the financial assistance from the Science Foundation Ireland Research Centre for AI-Driven Digital Content Technology, which allowed us to carry out this research. Finally, we would like to express our gratitude to the anonymous reviewers whose insightful critiques and recommendations have greatly raised the caliber of this work.

## 7. Bibliographical References

- Signon project. <https://signon-project.eu/>.
- Kshitij Bantupalli and Ying Xie. 2018. American sign language recognition using deep learning and computer vision. pages 4896–4899.

- Abul Abbas Barbhuiya, Ram Kumar Karsh, and Rahul Jain. 2021. Cnn based feature extraction and classification for sign language. *Multimedia Tools and Applications*, 80:3051–3069.
- Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. Isolated sign recognition from rgb video using pose flow and self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3441–3450.
- Neha Deshpande, Fabrizio Nunnari, and Eleftherios Avramidis. 2022. [Fine-tuning of convolutional neural networks for the recognition of facial expressions in sign language video samples](#). In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 29–38, Marseille, France. European Language Resources Association.
- Chen Chiung Hsieh, Dung Hua Liou, and David Lee. 2010. A real time hand gesture recognition system using motion history image. In *2010 2nd International Conference on Signal Processing Systems (ICSPS)*, volume 2. IEEE.
- Irish Deaf Society. Irish sign language. <https://www.irishdeafociety.ie/irish-sign-language/>.
- Hyunjun Jo, Yong Ho Na, and Jae Bok Song. 2017. Data augmentation using synthesized images for object detection. In *International Conference on Control, Automation and Systems*, pages 1035–1038.
- Doğan Keskin. 2023. [Synthetic data and data augmentation | by doğan keskin | medium](#).
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. [Deep learning](#). *Nature*, 521:436–444.
- Lorraine Leeson and John I. Saeed. 2012. *Irish Sign Language: A Cognitive Linguistic Approach*. Publisher Name.
- Dongxu Li. 2020. [Wlasl: A dataset for word-level american sign language](#).
- Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison.
- Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. [Findings of the first WMT shared task on sign language translation \(WMT-SLT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Irene Murtagh. 2021. The nature of verbs in sign languages: A role and reference grammar account of irish sign language verbs.
- Irene Murtagh, Víctor Ubieto Nogales, and Josep Blat. 2022. Sign language machine translation and the sign language lexicon: A linguistically informed approach. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 240–251.
- Marlon Oliveira, Houssein Chatbri, Ylva Ferstl, Mohamed Farouk, Suzanne Little, Noel E O'Connor, and Alistair S Sutherland. 2017. [A dataset for irish sign language recognition](#).
- Lorenzo Putzu, Luca Piras, and Giorgio Giacinto. 2020. Convolutional neural networks for relevance feedback in content based image retrieval: A content based image retrieval system that exploits convolutional neural networks both for feature extraction and for relevance feedback. *Multimedia Tools and Applications*, 79:26995–27021.
- RS Sabeenian, S Sai Bharathwaj, and M Mohamed Aadhil. 2020. Sign language recognition using deep learning and computer vision. *J Adv Res Dyn Control Syst*, 12(5 Special Issue):964–968.
- School of Linguistic, Speech and Communication Sciences. 2016. What is irish sign language and who uses it? <https://www.tcd.ie/slscs/faqs/irish-sign-language/>.
- Ankita Wadhawan and Parteek Kumar. 2020. Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications*, 32:7957–7968.
- Ryan Wong, Necati Cihan Camgöz, and Richard Bowden. 2022. Hierarchical i3d for sign spotting. In *European Conference on Computer Vision*, pages 243–255. Springer.
- Özge Yalçinkaya, Anil Atvar, and Pinar Duygulu. 2016. Hareket geçmişi görüntüsü yöntemi ile türkçe işaret dilini tanıma uygulaması. In *2016 24th Signal Processing and Communication Application Conference (SIU)*, pages 801–804. IEEE.

# Shedding Light on the Underexplored: Tackling the Minor Sign Language Research Topics

Jung-Ho Kim, Changyong Ko, Mathew Huerta-Enochian, and Seung Yong Ko

EQ4ALL

Nonhyeon-ro 76-gil 11, Gangnam-gu, Seoul, Republic of Korea

{stuartkim, ericko, mathew, stephenko}@eq4all.co.kr

## Abstract

In the past decade, sign language research has achieved remarkable results alongside advancements in deep learning. However, there is a disconnect between the outcomes of these research efforts and the actual use of sign language by signers. In this position paper, we reviewed sign language papers related to deep learning published in the last ten years to explore the reasons for this gap. We found many areas of research that are still underdeveloped, despite their linguistic importance. Based on an analysis of known corpora and methodologies, we identified the reasons for the lack of progress in these areas and propose directions for future research efforts.

**Keywords:** sign language research, underexplored research topics, sign language linguistics, communication methodologies

## 1. Introduction

We have seen many advances in sign language research with the introduction of deep learning. The most significant advances have been in recognition (Rastgoo et al., 2021a), translation (Kahlon and Singh, 2023), and generation (Rastgoo et al., 2021b). Despite severely limited resources, sign language research continues to make new advances every year.

Nevertheless, there are elements of sign language that are not studied despite being important linguistic elements (eye-gaze, topic, role-shifting, tensions, space allocation, depicting signs, buoys, etc.). These are important aspects of the language that are used in real life and should be studied if we want to make the results of sign language research practical.

In this position paper, we examine the elements of sign language linguistics, and investigate both actively researched areas and those that have received less attention. Furthermore, we propose why such studies are significant, discuss why certain studies have not been well-conducted, and what actions we should take to facilitate research in these areas.

## 2. Sign Language Linguistics

Sign languages employ visual-manual modalities, involving handshapes, movements, facial expressions, and body postures to convey meaning (Sutton-Spence and Woll, 1999; Valli and Lucas, 2000; Sandler and Lillo-Martin, 2006). This distinct mode of communication leads to unique linguistic structures, including phonology, morphology, syntax, and semantics, tailored to the visual-spatial

nature of sign languages. In this section, we delve into certain linguistic features that are more prominently highlighted in sign languages compared to spoken languages.

**Space and simultaneity** Sign languages are often referred to as spatial languages due to their inherent use of space to convey complex meanings. By exploiting the signing space with various articulators, signers can simultaneously present multiple pieces of information, a feature known as simultaneity (Geraci et al., 2008). For instance, signer use of buoys, which are handshapes or signs held in place to maintain a reference point or context while other signs are used to expand on other concepts, has been documented in various sign languages (Liddell, 2003; Tang et al., 2007). Simultaneous signs can represent actions, locations, or other descriptive information, allowing for a rich layering of language that is conveyed in a visually intuitive manner. This multi-layered approach to communication enables signers to present complex scenarios and narratives efficiently and effectively.

**Topicalization** Topicalization in sign languages involves the marking of a topic or the subject matter of a discourse at the beginning of a sentence or phrase, which is then followed by a comment or predicate about that topic. Friedman (1976); Aarons (1996) studied topicalization in American Sign Language (ASL) and Sze (2011) studied it in Hong Kong Sign Language (HKSL). This structure is often marked by specific non-manual signals such as raised eyebrows or a slight forward lean of the body, clearly distinguishing the topic from the rest of the discourse. This linguistic feature allows signers to structure their discourse in a way that



highlights the main points of discussion, making the communication clear and focused.

**Role-shifting** Role-shifting is a dynamic feature of sign languages where signers take on the roles of different characters in a narrative. Padden (1986) gave an early analysis of role-shifting in ASL and argued that its use is more than just play-acting. By physically shifting their body orientation, facial expressions, and gaze, signers can represent different perspectives and viewpoints within a story. Role-shifting adds depth to a narrative by allowing the signer to embody different characters, making utterances more engaging and easier to follow. This technique is not only a powerful storytelling tool but also a sophisticated linguistic mechanism for indicating changes in subject, object, and possessive relationships within a narrative.

**Phonology** Sign language phonology encompasses both spatial and temporal aspects of signing, a notable difference from spoken language phonology. Brentari (1998) explored both the simultaneity of ASL phonemes and asserted that movements are the most basic prosodic elements of ASL. Brentari (2011) later presented a thorough overview of phonology in sign languages, focusing on ASL but also drawing from studies on other sign languages. Much research has been devoted to exploring the building blocks of signing across different sign languages, usually focusing on articulator position, orientation, shape, and movement in the signing space. Temporal phonological features such as prosody and rhythm are also known to play a crucial role in most sign languages, adding layers of meaning and aiding in the conveyance of complex ideas and emotions. Cross-lingual variation has also been studied. For example, Tang et al. (2010) found that while eye blinks were used to mark certain intonational phrases in Japanese Sign Language (JSL), HKSL, Swiss German Sign Language (DSGS), and ASL, their use in JSL was unique out of these languages for blinks co-occurring with head nods rather than sign lengthening.

**Non-lexical expressions** Non-lexical expressions in sign languages encompass a range of communicative behaviors beyond the use of lexical signs, and include non-manual expressions (Valli and Lucas, 2000; Sandler and Lillo-Martin, 2006), depicting signs (Liddell, 2003; Cormier et al., 2012), and even gestures (Liddell and Metzger, 1998; Goldin-Meadow and Brentari, 2017). Non-manual expressions involve the use of facial expressions, body posture, and eye movements to convey meaning, mood, or grammatical information, adding

depth and nuance to the signed message. Depicting signs use handshapes and movements to represent objects, actions, or concepts, often providing visual and spatial information about the subject matter. Gestures, although not strictly part of the formal sign language lexicon, are incorporated into communication, offering a universal means of conveying ideas or emotions, sometimes transcending linguistic boundaries. Together, these elements enrich the expressive capacity of sign languages, allowing for a dynamic and multifaceted mode of communication.

### 3. Sign Language Research Topics

We examined research topics in sign language studies that applied deep learning and selected several representative topics, as can be seen in Figure 1. We also identified research topics with relatively few or no publications, despite being important linguistic aspects of sign language.

#### 3.1. Research Trends

We analyzed trends in sign language research from the past decade by reviewing a total of 544 papers from workshops, conferences, and journals in the fields of sign language, natural language processing, and computer vision. These papers were collected from the top twenty (by h5-index) publications in each of the following Google Scholar subcategories: Artificial Intelligence, Computational Linguistics, and Computer Vision & Pattern Recognition, in addition to selected papers from sign language workshops. The collection was restricted to works published between 2014 and early 2024. We categorized each paper by main topic and sub topic based on our interpretation of each paper's main focus. We provide our collection of relevant papers and paper topics through the digital repository link: <https://doi.org/10.5281/zenodo.10948417>.

**Recognition** Sign language recognition (SLR) involves automatically identifying handshapes, non-manual markers, fingerspellings, and glosses in video data and has seen the most active research (about 33% or 180 of 544 papers). Continuous and isolated SLRs are being advanced not only through improved feature extraction (He et al., 2016; Carreira and Zisserman, 2017; Xie et al., 2018) but also through new methods and applications such as better fusion of multiple input modalities (Chen et al., 2022), cross-frame feature trajectory analysis (Hu et al., 2023b), and knowledge distillation (Guo et al., 2023). Recently, sign spotting (Varol et al., 2022; Vázquez Enríquez et al., 2022) and sign segmentation (Woll et al., 2022; Moryossef et al., 2023) have

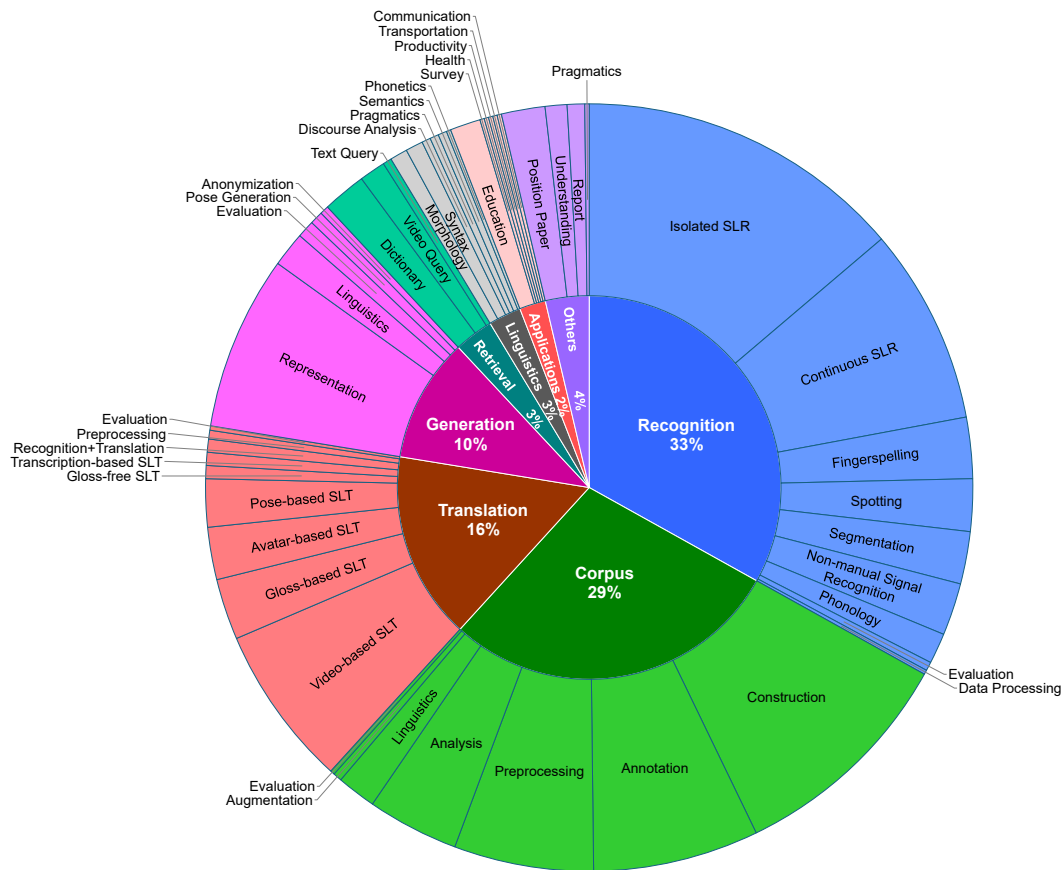


Figure 1: Research topics in sign languages. Tools for corpus, such as annotation and preprocessing, are included in each sub-topic depending on the purpose of the research.

emerged as prominent sub-topics.

**Translation** Sign language translation (SLT) is a task that translates between spoken language and sign language, or vice versa. Spoken language is represented through sound or text, and sign language is represented through gloss, skeletal pose, video (photo-realistic or avatar animation), or a notation system (usually SignWriting (Sutton, 2000) or HamNoSys (Hanke, 2004)). Recently, Gloss-free SLT, which translates sign language without the need for gloss supervision, has been actively researched (Yin et al., 2023; Lin et al., 2023; Zhou et al., 2023).

**Generation** Sign language generation (SLG) is the task of creating sign language poses or videos without translation<sup>1</sup>. Research has been conducted on a variety of topics, including the diversity of expressions (Kopf et al., 2023) and the anonymization of sign language users (Saunders et al., 2021; Xia et al., 2022). There have also been active propos-

<sup>1</sup>In this paper, we classify approaches that include both translation and generation as SLT and approaches that involve only sign language generation as SLG.

als for research aimed at reflecting sign language linguistics in the generation process (McDonald et al., 2014).

**Retrieval** While research on sign language dictionaries supporting word-based or handshape-based search has been active for some time, recent studies have focused on information retrieval through natural language queries in text (Duarte et al., 2022; Cheng et al., 2023) or video data (Sedmidubsky et al., 2018; De Coster and Dambre, 2023).

**Understanding** Although less researched than the other main topics, sign language understanding (SLU) has been explored in several ways. Recent studies have proposed methods for linguistically modeling sign languages (Mocialov et al., 2018; Hu et al., 2023a). An interesting development is the proposal of research on coreference resolution (Yin et al., 2021a) and a call to recognize SLU as a field within natural language processing (Yin et al., 2021b).

**Others** Sign language corpora have been crucial linguistic assets in sign language research for

an extended period, and corpora construction and analysis are areas that have received much focus. Additionally, applications and analysis of sign language in diverse areas such as health care, education, and communication have been proposed in academic papers. However, this paper focuses on deep learning-related research and does not extensively discuss these topics.

### 3.2. Underexplored Topics

Research in SLR, SLT, and SLG has advanced significantly, and yet some linguistic aspects of sign language modeling remain underexplored. This gap highlights a potential disconnect between research-generated output and actual signers' usage, underscoring the importance of incorporating sign language linguistics into future studies to ensure their authenticity and relevance. Below, we have listed research areas that, while being linguistically important in sign language, we believe are not being sufficiently researched.

**Elicitation methodologies** While elicitation methodologies have been studied extensively in traditional sign language corpora research and for spoken-language machine translation corpora, it has been mostly ignored in phrase-level sign language machine translation corpora, with few exceptions. [Matthes et al. \(2012\)](#) detailed how they developed tasks for capturing high-quality sign language utterances while still ensuring high overlap between multiple sign languages. [Huerta-Enochian et al. \(2022\)](#) compared several text-to-sign translation elicitation and revision methodologies and showed that text-based elicitation produced the least natural signing. Furthermore, we know that testing translation performance with back-translated data as the source language for spoken languages artificially inflates scores ([Zhang and Toral, 2019](#); [Graham et al., 2020](#)), but bias in development methodologies have not yet been explored for SLT.

**Pragmatics in SLT** SLT has now reached a level of maturity where it is poised to explore practicality in addition to novelty. To enhance the practical use of SLT, it is necessary to contemplate how to deliver sign language expressions from a pragmatic perspective. For instance, when translating and generating sign language, space should be used in concert with non-manual signals in order to generate easily-understandable translations. Recognition systems should be designed to handle a wide range of signs and integrate naturally with users without needing special gloves, cameras, or lights. Recently, [Fried et al. \(2023\)](#) called for increased focus on pragmatics for large language models

(LLMs), emphasizing the need for LLMs to adapt to the interlocutor. We suggest that this need is even greater for sign language modeling, given the crucial role of context in shaping how concepts are expressed.

**Depicting signs** Depicting signs are an area of research that is less frequently addressed in studies on SLR, SLT, and SLG. However, it is necessary to model depicting signs in each of these areas in order to approach the sign language representations actually used by signers. Since depicting signs are non-lexical expressions their use varies from person to person. There are many types of depicting signs, including the creation of gestures, the use of sign language to represent entities, and the description of situations through actions. Recent research on multi-modal large language models suggests new possibilities for exploring depicting signs. An important aspect of this research could be the representation of actions and relationships using one or both hands in sign language.

**Rhythm and tension** When generating sign language, the rhythm and stress of the signs are crucial elements in determining nuances. Similar to pragmatics, creating the appropriate sign language rhythm and stress according to the context will enable more natural sign language expressions and improve reception from the Deaf community.

**Others** There is a need for research on aspects that can be effectively used in sign language communication, such as topicalization and role-shifting. Moreover, translation between different sign languages could also present an intriguing area of study, potentially requiring methodologies distinct from those used in conventional translation. It is essential for research to more actively incorporate the history, culture, and linguistic aspects of sign language. There are also other areas in need of exploration, and we hope to see more proactive investigation of them in the future.

## 4. Challenges and Issues

We retrospectively examined existing studies with a focus on corpus and methodology challenges and explored how to resolve the issues identified.

### 4.1. Corpora

We examined a range of sign language corpora and summarized twenty-two commonly used corpora in [Table 1](#). Here we argue that the following considerations should be taken into account in the use, management, and further construction of sign language corpora.

Corpus	Language	Access	Video	Size	Channel	License
ASLG-PC12 (Othman and Jemni, 2012)	ASL	open	N	77M (24M)	single	CC BY-NC 4.0
ATIS (Bungeroth et al., 2008)	multilingual (DGS,ISL,SASL)	restricted	Y	595 (-)	multi	CC BY-NC 4.0
AUSLAN (Johnston, 2009)	Auslan	restricted	Y	- (-)	multi	CC BY-NC-ND 4.0
BSL Corpus (Schembri et al., 2017)	BSL	open(partial) / academic	Y	- (14,754)	multi	custom
BOBSL (Albanie et al., 2021)	BSL	restricted	Y	1.2M (-)	multi	custom
CONTENT4ALL (SWISSTXT-WEATHER) (Camgöz et al., 2021)	DSGS	restricted	Y	811 (-)	single	CC BY-NC-SA 4.0
CONTENT4ALL (SWISSTXT-NEWS) (Camgöz et al., 2021)	DSGS	restricted	Y	6,031 (-)	single	CC BY-NC-SA 4.0
CONTENT4ALL (VRT-NEWS) (Camgöz et al., 2021)	VGT	restricted	Y	7,174 (-)	single	CC BY-NC-SA 4.0
Corpus NGT (Crasborn and Zwitserlood, 2008)	NGT	open(partial) / restricted	Y	- (490)	multi	CC BY-NC-SA 4.0
CSL-Daily (Zhou et al., 2021)	CSL	academic	Y	20,654 (-)	single	custom
Dicta Sign (Matthes et al., 2012)	BSL, DGS GSL, LSF	academic / restricted	Y	- (-)	single	-
KETI (Ko et al., 2019)	KSL	restricted	Y	2,940 (-)	single	-
KRSL-OnlineSchool (Mukushev et al., 2022)	KRSL	restricted	Y	1M (-)	single	-
NCSLGR (Neidle and Vogler, 2012)	ASL	open	Y	1,887 (1,874)	multi	custom
NIASL2021 (Huerta-Enochian et al., 2022)	KSL	open(domestic)	Y	201,026 (180,892)	multi	custom
DGS Corpus (Konrad et al., 2020)	DGS	open(partial) / restricted	Y	- (63,922)	multi	custom
RWTH-BOSTON-104 (Dreuw et al., 2007)	ASL	open	Y	201 (201)	single	-
RWTH-PHOENIX-WEATHER-2014-T (Camgoz et al., 2018)	DGS	open	Y	8,257 (8,257)	single	CC BY-NC-SA 3.0
SignBank <sup>◇</sup>	multilingual	open	N	- (29,035)	multi	-
STS Corpus (Öqvist et al., 2020)	SSL	open(web-access) / registered	Y	- (-)	multi	CC By-NC-SA 4.0
RWTH-PHOENIX-WEATHER 2014 (Forster et al., 2014)	DGS	open	Y	6,861 (6,841)	single	CC BY-NC-SA 3.0
How2Sign (Cardoso Duarte et al., 2021)	ASL	open(w/o gloss)	Y	35,191 (-)	single	CC BY-NC 4.0

◇: <https://www.signbank.org/signpuddle/>, accessed on February 23, 2024

Table 1: Summary of reviewed corpora. We limited reporting to *sentence-level* data. **Access:** *open*, *registered* (available with registration), *academic* (available for non-commercial research or academia), and *restricted* (available only with explicit permission). We report multiple levels as applicable. **Size:** The reported sentence-level instance count and our calculated open access count, if available. Every effort was made to report correct sizes for open access data, but there may be some deviation based on access method. **Channel:** Data is categorized based on the presence of annotations for separate hands or for non-manual signals, regardless of the existence of multiple tiers. **License:** The current corpus license. Note that licenses may differ from those reported in original research or from software licenses.

**Data format** The central challenge to choosing a data annotation format is that sign representation fidelity is inversely related to representation simplicity. In other words, simple representations like glosses cannot adequately represent the nuances of multiple signed instances while more informative

representations like sign writing or even pose data are not easy to work with. This leads to variations in data and glossing formats across corpora, which in turn requires significant additional preprocessing before corpora can be used for training (De Sisto et al., 2022). Recently, there has been more inter-



est in rectifying this issue as can be seen in the proposed rectification of annotations from the easier project [Kopf et al. \(2022\)](#) and in [Schulder et al. \(2023\)](#) proposal of the sign language interchange format. While rectifying these differences between corpora is a good and necessary solution, using more unified annotation conventions for future corpus projects will be immensely helpful.

**Data availability** Though many corpora have been released for sign language research, collection and use of potential corpora is complicated by missing data links, mixed access levels, and custom licenses. Notably, some corpora were publicly available at the time of publication but are no longer accessible.

**Commercial-friendly data** Only two of the corpora we reviewed explicitly support commercial applications: the partially open release of the BSL Corpus and NIASL2021 (which is currently limited to users in Korea). In addition, five of the corpora do not include specific licensing information, introducing legal risks if used. The vast majority of corpora use derivatives of CC BY-NC or custom licenses that designate corpora for research purposes only. To encourage research from industry as well as academia, it may be necessary to reflect an incentive mechanism for data disclosure. However, in this case, ethical considerations such as re-obtaining consent from contributors due to a changing release policy and data anonymization should also be taken into account.

**Signing quality** Sign language corpora for machine learning show much variation in terms of signing quality. One major factor in this is the range of elicitation and collection methodologies. Some corpora feature only spontaneous utterances on open-ended topics, some corpora focus elicitation to specific tasks, and many corpora use either real-time interpretation or pre-translated utterances. We are not advocating against using specific corpora. On the contrary, given the small size of available data, utilizing existing corpora as much as possible—including corpora containing non-spontaneous signing—is necessary. While the effectiveness of high-quality training data is undisputed, lower-quality data is often utilized for pretraining, contrastive training, and other novel approaches. A key challenge moving forward will be to better classify signing data by recommended use and to improve elicitation techniques in general.

## 4.2. Methodologies

**Text-to-sign translation** There is a growing interest in direct pose- and video-predicting models, likely due to the lower annotation burden and the

appeal of end-to-end solutions. While visualization of single-channel gloss data is limited, it still has significant value for identifying bias, data balance issues, linguistic insights, for researchers invested in procedural generation, and in hybrid approaches. Similarly, high-cost annotations like multi-channel glosses and notation systems offer the possibility of higher fidelity translations in specific domains. While we agree that the potential of end-to-end solutions are the most promising in the long-term, we urge the community to keep prioritizing multiple data modalities given the continued need for both short-term and long-term solutions.

**Modeling non-lexical signs** Procedural generation of non-lexical signs from gloss annotations is extremely challenging. High-detail annotations like multi-channel glosses, AZee, HamNoSys, Sign-Writing, and other phonetic annotations provide additional possibilities for non-lexical sign generation. While end-to-end solutions should be able to produce non-lexical signs, hybrid approaches like the one explored by [Saunders et al. \(2022\)](#) are likely more realistic in the short-term.

Non-lexical sign recognition is also an area that may likely benefit from novel approaches, particularly by delving into the intricacies of sign language. Effective recognition of non-lexical signs may involve understanding sign language morphemes, identifying what entities the handshapes represent, or even interpreting the intent behind gestures. This deeper comprehension could lead to more effective communication aids for the deaf and hard of hearing, by not just recognizing signs as whole units but understanding their component parts and the meanings they convey in different contexts.

**SLT automatic evaluation** While traditional machine translation metrics (BLEU, Rouge, etc.) can be applied to simple gloss translations, there is no definitive metrics for the many other text→sign output representations. Recently applied and proposed metrics include SignBLEU ([Kim et al., 2024](#)) for multi-channel gloss prediction; BLEU, chrF2++, and mean absolute error metrics for Formal Sign-Writing ASCII ([Jiang et al., 2023](#)); and SLR pose classification accuracy ([Xiao et al., 2020](#)) and Fréchet Gesture Distance ([Yoon et al., 2020](#)) for pose prediction. As a community, we need to continue researching and improving potential metrics.

**Human evaluation accessibility** A hugely influential factor in the creation, evaluation, and curation of text-based data has been Amazon’s Mechanical Turk (MTurk). While MTurk can be used for sign language data, finding highly-specialized participants through MTurk is a known challenge ([Chandler and Shapiro, 2016](#)). Furthermore, due to

communication barriers and ethics board requirements, most human evaluation of sign language is conducted locally and limited to the local sign language. This means that most machine learning research is human-validated on a single sign language or none at all. We see the need for better international cooperation, preferably as an official network, devoted to ensuring high-quality human evaluation for sign language applications.

## 5. Possible Directions

In the previous section, we placed as much importance on the practicality of sign language research outcomes as on improving performance. Here, we provide insights into areas that we think should receive increased focus in future research.

**Additional annotation of existing corpora** We have observed that understanding the position and direction in sign language plays a crucial role in comprehending its syntax. Therefore, transcribing this information, either automatically or manually, and applying it to SLR, SLT, and SLG models is essential.

**Elicitation methodologies** In light of the data scarcity problem in SLR, SLT, and SLG, the quality of signing data is of increased importance, and there are several urgent research directions to be explored. Data for specific translation applications usually requires highly-structured translations from existing spoken-language text. However, improving the quality of text-to-sign translations while ensuring high content fidelity is an open problem. As mentioned in section 3, traditional corpora research suggests using language-neutral elicitation materials, but applying such media to translation of specific phrases needs more exploration. In order to avoid bias, we need to research proper methodologies for sign language translation train and test set construction.

**Pragmatics** Pragmatics in sign language explores how language functions within social contexts and interactions. To address this, a deep learning model methodology is essential—one that not only minimizes ambiguity but also ensures communication objectives are met through word choice, spatial utilization, and the use of non-manual expressions. Establishing a clear evaluation framework is equally crucial to assess a model's overall effectiveness in enhancing clarity and communication efficiency. [Fried et al. \(2023\)](#) proposed how to model pragmatics with large language models to achieve these communication goals for all natural languages.

**Non-lexical signs** Effective modeling of non-lexical signs will require novel solutions, and we expect that many potential solutions will be found in linguistic insights. For example, [Taub \(2001\)](#) first proposed the analogue-building model process which is comprised of three steps (image selection, schematization, and encoding), and subsequent studies ([Emmorey, 2014](#); [Nordheimer et al., 2024](#)) have built on and applied this model. We see the potential of this method applied to SLT through an approach using knowledge distillation and representation learning as a way to train entity translation in a generalizable way.

**Hate speech** The exploration of hate speech in sign language research is essential for the development of protective measures and educational tools that can help safeguard communities from discrimination and abuse. The nuanced gestures and expressions unique to sign languages can convey complex emotions and intentions, making it vital to understand how hate speech manifests in these modes of communication. Consequently, building comprehensive corpora that capture the breadth of sign language expressions, including those that could be considered hate speech, is imperative. These corpora will not only facilitate the identification and mitigation of hate speech within sign language communication but also contribute to the broader efforts of promoting digital safety and inclusiveness for all, regardless of mode of communication.

**Deaf involvement** Currently, Deaf involvement in sign language machine learning research is largely limited to participation in corpora construction and annotation and in human evaluation of developing technologies. Limited Deaf involvement in research means that hearing-centric views may grow unchecked and we risk losing sight of meaningful research objectives. On the other hand, increased involvement will provide insights to which non-native signers do not have access and ensure that we work towards developing solutions that the community can actually use.

## 6. Concluding Remarks

We have explored areas within sign language research that have not been well addressed. We also examined and proposed directions for future research in these areas. We argue that future sign language studies should be more closely connected with sign language linguistics and reconsider their practicality. We hope that by doing so, research outcomes will be more readily accepted in Deaf communities. Not all research topics could be covered in this paper, and as research progresses,

downstream tasks of NLP that are currently underexplored for sign language, including summarization, question answering, and language modeling, will likely receive more attention.

## Acknowledgment

This work was partly supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00010, Development of Korean sign language translation service technology for the deaf in medical environment) and Artificial intelligence industrial convergence cluster development project funded by the Korean government (MSIT) & Gwangju Metropolitan City (No. BA00000797, LLM-based sign language translation for weather forecasts).

## Sign Language Abbreviations

**ASL** American Sign Language  
**Auslan** Australian Sign Language  
**BSL** British Sign Language  
**CSL** Chinese Sign Language  
**DGS** German Sign Language  
**DSGS** Swiss-German Sign Language  
**GSL** Greek Sign Language  
**ISL** Irish Sign Language  
**KRSL** Kazakh–Russian Sign Language  
**KSL** Korean Sign Language  
**LSF** French Sign Language  
**NGT** Dutch Sign Language  
**SASL** South African Sign Language  
**SSL** Swedish Sign Language  
**VGT** Flemish Sign Language

## Bibliographical References

Debra Aarons. 1996. Topics and topicalization in american sign language. *Stellenbosch Papers in Linguistics*, 30(1):65–106.

Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, et al. 2021. Bbc-oxford british sign language dataset. *arXiv preprint arXiv:2111.03635*.

Diane Brentari. 1998. *A prosodic model of sign language phonology*. Mit Press.

Diane Brentari. 2011. Sign language phonology. *The handbook of phonological theory*, pages 691–721.

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. [The ATIS sign language corpus](#). In *6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2943–2946, Marrakech, Morocco. European Language Resources Association (ELRA).

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. [Content4all open research sign language translation datasets](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.

Amanda Cardoso Duarte, Shruti Palaskar, Lucas Ventura Ripol, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres Viñals, and Xavier Giró Nieto. 2021. [How2sign: A large-scale multimodal dataset for continuous american sign language](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition: Virtual, 19-25 June 2021: proceedings*, pages 2734–2743. Institute of Electrical and Electronics Engineers (IEEE).

Joao Carreira and Andrew Zisserman. 2017. [Quo vadis, action recognition? a new model and the kinetics dataset](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jesse Chandler and Danielle Shapiro. 2016. Conducting clinical research using crowdsourced convenience samples. *Annual review of clinical psychology*, 12:53–81.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022. Two-stream network for sign language recognition and translation. *Advances in Neural Information Processing Systems*, 35:17043–17056.

Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. 2023. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19016–19026.

Kearsy Cormier, David Quinto-Pozos, Zed Sevcikova, and Adam Schembri. 2012. Lexicalisation and de-lexicalisation processes in sign languages: Comparing depicting constructions and

- viewpoint gestures. *Language & communication*, 32(4):329–348.
- Onno Crasborn and Inge Zwisserlood. 2008. [The Corpus NGT: an online corpus for professionals and laymen](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 44–49, Marrakech, Morocco. European Language Resources Association (ELRA).
- Mathieu De Coster and Joni Dambre. 2023. Querying a sign language dictionary with videos using dense vector search. In *Proceedings of the 8th International Workshop on Sign Language Translation and Avatar Technology*. IEEE.
- Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. [Challenges with sign language datasets for sign language recognition and translation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.
- Philippe Dreuw, David Rybach, Thomas Dese-laers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. *hand*, 60:80.
- Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. 2022. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14094–14104.
- Karen Emmorey. 2014. Iconicity as structure mapping. *Philosophical transactions of the Royal Society B: Biological sciences*, 369(1651):20130301.
- Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus rwth-phoenix-weather. In *LREC*, pages 1911–1916.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2023. [Pragmatics in language grounding: Phenomena, tasks, and modeling approaches](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12619–12640, Singapore. Association for Computational Linguistics.
- Lynn A Friedman. 1976. The manifestation of subject, object, and topic in american sign language. *Word Order and Word Order Change*, pages 940–961.
- Carlo Geraci, Marta Gozzi, Costanza Papagno, and Carlo Cecchetto. 2008. How grammar can cope with limited short-term memory: Simultaneity and seriality in sign languages. *Cognition*, 106(2):780–804.
- Susan Goldin-Meadow and Diane Brentari. 2017. Gesture, sign, and language: The coming of age of sign language and gesture studies. *Behavioral and brain sciences*, 40:e46.
- Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. [Statistical power and translationese in machine translation evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.
- Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Shengyong Chen. 2023. Distilling cross-temporal contexts for continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10771–10780.
- Thomas Hanke. 2004. Hamnosys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023a. [Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11221–11239.
- Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. 2023b. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2529–2539.
- Mathew Huerta-Enochian, Du Hui Lee, Hye Jin Myung, Kang Suk Byun, and Jun Woo Lee. 2022. Kosign sign language translation project: Introducing the niasl2021 dataset. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 59–66.



- Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. [Machine translation between spoken languages and signed languages represented in SignWriting](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.
- Trevor Johnston. 2009. Creating a corpus of auslan within an australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages*.
- Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1):1–35.
- Jung-Ho Kim, Mathew Huerta-Enochian, Changyong Ko, and Du Hui Lee. 2024. Signbleu: Automatic evaluation of multi-channel sign language translation. Accepted.
- Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683.
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Wörseck, Oliver Böse, Elena Jahn, and Marc Schuler. 2020. [Meine dgs – annotiert. öffentliches korpus der deutschen gebärdensprache, 3. release / my dgs – annotated. public corpus of german sign language, 3rd release](#).
- Maria Kopf, Rehana Omardeen, and Davy Van Landuyt. 2023. Representation matters: The case for diversifying sign language avatars. In *Proceedings of the 8th International Workshop on Sign Language Translation and Avatar Technology*. IEEE.
- Maria Kopf, Marc Schuler, Thomas Hanke, and Sam Bigeard. 2022. [Specification for the harmonization of sign language annotations](#). Project Deliverable EASIER D6.2, EASIER project, IDGS, Hamburg University, Hamburg, Germany.
- Scott K Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.
- Scott K Liddell and Melanie Metzger. 1998. Gesture in sign language discourse. *Journal of pragmatics*, 30(6):657–697.
- Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. [Gloss-free end-to-end sign language translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.
- Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Wörseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. [Dicta-sign-building a multilingual sign language corpus](#). In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Satellite Workshop to the eighth International Conference on Language Resources and Evaluation (LREC-2012)*.
- John C. McDonald, Rosalee Wolfe, Robyn Moncrief, and Souad Baowidan. 2014. [Analysis for synthesis: Investigating corpora for supporting the automatic generation of role shift](#). In *Proceedings of the LREC2014 6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel*, pages 117–122, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Boris Mocialov, Helen Hastie, and Graham Turner. 2018. [Transfer learning for British Sign Language modelling](#). In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 101–110, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023. [Linguistically motivated sign language segmentation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.
- Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. [Data augmentation for sign language gloss translation](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.
- Medet Mukushev, Aigerim Kydyrbekova, Vadim Kimmelman, and Anara Sandygulova. 2022. Towards large vocabulary kazakh-russian sign language dataset: Krsl-onlineschool. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 154–158.

- Carol Neidle and Christian Vogler. 2012. A new web interface to facilitate access to corpora: Development of the asllrp data access interface (dai). In *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC*, volume 3. Citeseer.
- Swetlana Nordheimer, Allison Marlow, and Janina Scholtz. 2024. Fostering mathematical creativity and talents with mathematical problems and competitions in german sign language. *The 13th IMCGC Bloemfontein*.
- Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. [STS-korpus: A sign language web corpus tool for teaching and public use](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).
- Achraf Othman and Mohamed Jemni. 2012. English-asl gloss parallel corpus 2012: Asl-gpc12. In *sign-lang@ LREC 2012*, pages 151–154. European Language Resources Association (ELRA).
- Carol Padden. 1986. Verbs and role-shifting in american sign language. In *Proceedings of the fourth national symposium on sign language research and teaching*, volume 44, page 57. National Association of the Deaf Silver Spring, MD.
- Pierre Poitier, Jérôme Fink, and Benoît Frénay. 2024. Towards better transition modeling in recurrent neural networks: The case of sign language tokenization. *Neurocomputing*, 567:127018.
- Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021a. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.
- Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021b. Sign language production: A review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3451–3461.
- Wendy Sandler and Diane Carolyn Lillo-Martin. 2006. *Sign language and linguistic universals*. Cambridge University Press.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. [Anonymsign: Novel human appearance synthesis for sign language video anonymisation](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151.
- Adam C. Schembri, Jordan B Fenlon, Ramas Rentelis, and Kearsy Cormier. 2017. [British sign language corpus project: A corpus of digital video data and annotations of british sign language 2008-2017](#).
- Marc Schulder, Sam Bigeard, Thomas Hanke, and Maria Kopf. 2023. [The sign language interchange format: Harmonising sign language datasets for computational processing](#). In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops: Sign Language Translation and Avatar Technology (SLTAT 2023)*, Rhodes, Greece. IEEE.
- Jan Sedmidubsky, Petr Elias, and Pavel Zezula. 2018. Effective and efficient similarity searching in motion capture data. *Multimedia Tools and Applications*, 77:12073–12094.
- Valerie Sutton. 2000. Signwriting. *Deaf Action Committee (DAC) for Sign Writing*.
- Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.
- Felix Sze. 2011. Nonmanual markings for topic constructions in hong kong sign language. *Sign Language & Linguistics*, 14(1):115–147.
- Gladys Tang, Diane Brentari, Carolina González, and Felix Sze. 2010. *Crosslinguistic variation in prosodic cues*. na.
- Gladys Tang, Felix Sze, Scholastica Lam, et al. 2007. Acquisition of simultaneous constructions by deaf children of hong kong sign language. *Simultaneity in signed languages*, pages 283–316.
- Sarah F Taub. 2001. *Language from the body: Iconicity and metaphor in American Sign Language*. Cambridge University Press.

- Clayton Valli and Ceil Lucas. 2000. *Linguistics of American sign language: An introduction*. Galaudet University Press.
- Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. Scaling up sign spotting through sign language dictionaries. *International Journal of Computer Vision*, 130(6):1416–1439.
- Manuel Vázquez Enríquez, José L Alba Castro, Laura Docio Fernandez, Julio CS Jacques Junior, and Sergio Escalera. 2022. Eccv 2022 sign spotting challenge: Dataset, design and results. In *European Conference on Computer Vision*, pages 225–242.
- Bencie Woll, Neil Fox, and Kearsy Cormier. 2022. [Segmentation of signs for research purposes: Comparing humans and machines](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 198–201, Marseille, France. European Language Resources Association (ELRA).
- Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitris Metaxas. 2022. [Sign language video anonymization](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 202–211, Marseille, France. European Language Resources Association (ELRA).
- Qinkun Xiao, Minying Qin, and Yuting Yin. 2020. Skeleton-based chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, 125:41–55.
- Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562.
- Kayo Yin, Kenneth DeHaan, and Malihe Alikhani. 2021a. [Signed coreference resolution](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4950–4961, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021b. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.
- Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16.
- Mike Zhang and Antonio Toral. 2019. [The effect of translationese in machine translation test sets](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.
- Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.
- Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.

# Headshakes in NGT: Relation between Phonetic Properties & Linguistic Functions

Vadim Kimmelman<sup>1</sup> , Marloes Oomen<sup>2</sup> , Roland Pfau<sup>2</sup> 

<sup>1</sup>University of Bergen, <sup>2</sup>University of Amsterdam

<sup>1</sup>Bergen, Norway, <sup>2</sup>Amsterdam, the Netherlands

vadim.kimmelman@uib.no, {M.Oomen2, R.Pfau}@uva.nl

## Abstract

Non-manual markers (such as facial expressions and head movements) have been shown to fulfil a wide range of grammatical functions across sign languages (Pfau and Quer, 2010). One nonmanual marker that is very wide-spread is headshake used to express negation (Oomen and Pfau, 2017). While negation and headshake have been studied for a variety of sign languages, phonetic/kinematic research on headshake has been mostly absent. In this paper, we conduct a phonetic analysis of headshake in Sign Language of the Netherlands using a Computer Vision solution, namely OpenFace (Baltrusaitis et al., 2018). We specifically analyze whether linguistic properties of headshake (e.g. spreading and the type of signs co-occurring with the headshake) influence its phonetic form.

**Keywords:** headshake, negation, OpenFace, Sign Language of the Netherlands

## 1. Introduction

Non-manual markers (such as facial expressions and head movements) have been shown to fulfil a wide range of grammatical functions across sign languages (SLs) (Pfau and Quer, 2010; Wilbur, 2021). Moreover, it has been observed that one and the same non-manual may display different properties depending on whether it is used grammatically or as a co-speech gesture. Zooming in on grammatical uses, a certain non-manual may also fulfil various grammatical functions within a given SL (e.g., brow raise; Wilbur and Patschke 1999). Yet, to date, very few studies have addressed the question whether subtle phonetic differences might also distinguish between various functions of a multifunctional marker. In the present study, we address this question for the headshake, as used in SL of the Netherlands (NGT), using naturalistic corpus data and Computer Vision processing technology.

## 2. Background

### 2.1. Manual and Non-Manual Negation in Sign Languages

The expression of clausal negation is one of the best-studied domains of grammar for sign languages: negation has been described for a considerable number of both urban and rural SLs, there are some comparative studies available (Pfau and Quer, 2002; Zeshan, 2004, 2006; Pfau, 2016), and handbook chapters provide convenient overviews of the phenomenon (Quer, 2012; Gökgöz, 2021). These studies reveal that all SLs studied to date employ manual and non-manual markers of nega-

tion, i.e., negative elements that are manually expressed, and head movements or other non-manual elements that are articulated simultaneously with (strings of) signs. However, the ways in which manual and non-manual negators interact within a clause has been shown to be subject to interesting cross-linguistic variation.

On the one hand, there are SLs in which the use of a manual negator is obligatory; this negator is then commonly, but not obligatorily, accompanied by a headshake – or, in some geographical areas, by a backward head tilt (e.g., in Turkish SL; Makaroğlu 2021).<sup>1</sup> However, the non-manual does not usually spread onto neighboring constituents. Such SLs are labeled *manual dominant* SLs (Zeshan 2004).

An example from Inuit SL is provided in (1a); here, the manual particle NEG occupies a clause-final position and is accompanied by a headshake ('hs'). The example in (1b), without NEG, is ungrammatical, irrespective of the scope of the headshake (Schuit, 2013, 48,50).

- (1) a. WOLVERINE EAT  $\overset{\text{hs}}{\text{NEG}}$   
'I don't eat wolverine.'  
b. \*POLAR.BEAR SEE  $\overset{\text{hs}}{\text{NEG}}$   
'I didn't see a polar bear.'

In contrast, in other SLs, it is possible – and actually common – to encode clausal negation by means of only a headshake. Manual negative particles do exist but their use is optional. Moreover,

<sup>1</sup>Further non-manual markers of negation have been described in the literature, e.g., a 'negative facial expression' (Yang and Fischer 2002 for Chinese SL) and tongue protrusion (Lutzenberger et al. 2022 for Kata Kolok).



in such *non-manual dominant* SLs, the headshake may spread over parts of the clause, e.g., the verb or the entire verb phrase (Zeshan 2004). NGT has been shown to belong to this typological group. We will provide examples in the next section.

It remains to be emphasized that recent studies suggest that the dichotomy (originally put forward in Zeshan 2004) may not be sufficient. Some SLs present us with a hybrid picture in that they require the use of a manual negator, but still spreading of the headshake beyond the negative particle is possible (e.g., Rudnev and Kuznetsova 2021 for Russian SL; Pfau et al. 2022 for Georgian SL).

An important assumption underlying both the above-mentioned investigations and the present study is that the headshake, as used in these sign languages, is indeed a grammatical marker. Of course, headshakes are also commonly used as co-speech gesture in spoken interactions (e.g., Kendon 2002; Harrison 2014). However, the fact that the use and distribution of headshake across sign languages has been shown to be subject to language-specific constraints suggests that it is not a mere gesture but rather functions as a linguistic non-manual marker (see Pfau 2015 on the grammaticalization of headshake). This does not exclude the possibility that in a given sign language, the headshake is not (yet) grammaticalized (as has been argued for Australian Sign Language by Johnston 2018).

## 2.2. Negation and Headshake in NGT

Oomen and Pfau (2017) present a corpus-based study on the realization of standard negation in NGT. The study is based on the analysis of 120 negative clauses, including clauses with non-verbal predicates, identified in the Corpus NGT (see Section 3.1 for details). Note that Oomen and Pfau additionally annotated clauses containing neg-words, such as NOTHING or NEVER, as well as clauses containing negative modals; however, these cases were excluded from the analysis as they were not considered standard negation.<sup>2</sup> The attested patterns clearly show that NGT can be classified as a non-manual dominant SL – thus confirming earlier observations by Coerts (1992) and Brunelli (2011): 47 clauses (39.2%) contain the negative particle NOT (2a, 2b) while 70 clauses (58.3%) are negated by headshake only (2c) (three clauses involve negative concord and were excluded). For the former group, they further observe that NOT may either follow the verb (which often is also the clause-final position), as in (2a), or precede the verb phrase (2b).

<sup>2</sup>For a general overview of NGT negation, see Klomp et al. in press; for negative concord in NGT, see Van Boven et al. 2023.

- (2) a. IX<sub>1</sub> POINT  $\overline{\text{UNDERSTAND NOT}}^{\text{hs}}$   
 'I don't understand/get the point.' [390-S019-00:53]
- b. IX<sub>1</sub> ACTUALLY  $\overline{\text{NOT LEARN}}^{\text{hs}}$   
 'I'm not going to learn (it).' [065-S006-01:25]
- c. IX<sub>3</sub> SELF BASIS  $\overline{\text{STRONG ENOUGH}}^{\text{hs}}$  IX<sub>3</sub>  
 'Their basis isn't strong enough.' [386-S019-00:22]

As for the headshake, Oomen and Pfau (2017) report the following observations:

- When NOT is present, it is always accompanied by a headshake.
- Predicates are accompanied by headshake in 94% of all negative clauses.
- Objects, when present, may or may not (2a) be accompanied by headshake, no matter whether they are nominal or pronominal.
- Subjects are only accompanied by a headshake if they are pronominal (only one exception in their dataset).
- Elements that follow the verb, like pointing signs (2c) or PALM-UP (3) may be accompanied by headshake.

Based on this distribution, they claim that in NGT, the headshake may fulfil up to three different linguistic functions within a single clause, as shown in (3): (i) for the manual negator, it is *lexically* specified (hs<sub>L</sub>); (ii) when accompanying the predicate, it functions as a simultaneous *morphological* affix (hs<sub>M</sub>); and (iii) it may optionally spread over additional elements in the clause for *prosodic* purposes (hs<sub>P</sub>). The claim regarding prosodic spreading is motivated by the observation that prosodically light elements such as pronominal subjects and clause-final pointing signs and PALM-UP (3) are commonly accompanied by headshake. As indicated in (3), the headshake is not interrupted but rather is articulated as a continuous contour across multiple manual signs (Oomen et al., 2018, 45).

- (3) DEAF SELF IX<sub>3</sub>  $\overline{\text{HAVE.PROBLEM NOT PU}}^{\text{hs}_M \text{hs}_L \text{hs}_P}$   
 'The deaf themselves don't have a problem (with it).' [387-S019-01:26]

## 2.3. Quantitative Analysis of Headshake

Not a lot of quantitative research on headshake in sign languages exists, to our knowledge. Harmon (2017) reports that ASL uses two types of headshake that differ in phonetic characteristics,

but does not provide specific quantitative results. Chizhikova and Kimmelman (2022) previously conducted a study of negative headshake in Russian SL (RSL), using OpenFace (Baltrusaitis et al., 2018) to measure phonetic properties of headshake; the current study is partially an application of the same approach to NGT. In the quantitative analysis, the authors analyzed 68 instances of negative headshake from the online corpus of RSL (Burkova, 2015). For each instance, they calculated the number of peaks (reflecting the number of turns of the head), frequency and maximal amplitude, and the average measures they found in the data. Chizhikova and Kimmelman (2022) show that these measures do not correlate with the type of manual negative sign that is accompanied by the headshake.

It is important to note that RSL is quite different from NGT in terms of negative headshake, as it is clearly a manual-dominant language in the domain of negation; Chizhikova and Kimmelman (2022) found that headshake is clearly optional in negative sentences (only 28% of such sentences in the corpus had headshake). Furthermore, following the general typological trend for manual-dominant languages, while spreading of the headshake is possible (Rudnev and Kuznetsova, 2021), it is clearly rare (13% of the cases).

Given that NGT is a non-manual dominant language in which the headshake tends to spread beyond the manual negator (if present), it is reasonable to expect that the phonetic properties of headshake in NGT may be substantially different than in RSL. We therefore aim to explore possible correlations between the linguistic functions of headshake and its phonetics properties in NGT. Following Chizhikova and Kimmelman (2022), the properties we are analyzing include number of peaks, frequency, and maximal amplitude. We expect that the measures for these properties will differ depending on the predicted linguistic function of headshake, in line with Oomen and Pfau (2017). More precisely, we expect to find differences between lexical, morphological and prosodic spreading in terms of phonetic characteristics of the headshake.

### 3. Methodology

#### 3.1. The dataset

For our study, we used the annotated dataset compiled by Oomen and Pfau (2017). The authors analyzed 35 video clips (amounting to approx. 95 minutes of data) from the Corpus NGT, which includes (partially) annotated video files of stories and conversations between deaf native signers of NGT (Crasborn et al., 2008). The selected videos involve 22 signers (14 female, 8 male), all from

the Groningen region, with an age range between 18–50 years. As mentioned before, Oomen and Pfau analyzed 120 negative clauses from these videos, all involving standard negation. However, in contrast to them, we also include three instances of negative concord, as well as negated clauses involving negative modals (N = 21), the neg-words NOTHING and NEVER (N = 39), or the negative completive NOT-YET (N = 5), which they identified in the original data set but did not analyze further. Moreover, we coincidentally spotted one negated example that had apparently been overlooked by Oomen and Pfau. This leaves us with 220 instances of headshake for analysis.

#### 3.2. Annotation

All 35 videos had previously been annotated in ELAN (Crasborn and Sloetjes, 2008) for manual signs (right and left hand on separate tiers for both signers) by the Corpus NGT team; most of the videos additionally included Dutch translations. Oomen and Pfau (2017) added a tier *Headshake*, on which they annotated the presence and the scope of the headshake. For the present study, we reviewed the annotations on the *Headshake* tier and made a few corrections. Furthermore, two additional tiers were created:

- *ManualNegation*: On this tier, we specified the type of manual negative sign(s) in the clause, if present. Four annotation values were distinguished – ‘Neg’ (for the standard clause negator), ‘Neg.Mod’ (for negative modals), ‘Neg.Word’ (for neg-words), and ‘Neg.Comp’ (for the negative completive NOT-YET). This tier allowed us to differentiate between clauses with standard negation and clauses involving other types of negation. Clauses with standard negation are those that (a) include the annotation ‘Neg’, or (b) do not include an annotation on this tier (but involve headshake only). Almost all clauses that include a manual negative sign also include a headshake (annotated on the main *Headshake* tier), although there are a handful of exceptions, typically involving manual negative signs other than the basic clause negator.
- *HeadshakeType*: On this tier, we annotated the linguistic function of a headshake, taking the claims made by Oomen and Pfau (2017) as point of departure; three annotation values were distinguished – ‘Lex’ for lexically specified headshake (accompanying negative signs), ‘Morph’ for morphological headshake (accompanying the predicate), and ‘Pros’ for prosodic headshake (accompanying all other signs in a clause). The annotations were aligned with the

scope of the annotations for the relevant manual signs, thus excluding the transition periods between signs.

### 3.3. Computer Vision Processing

We extracted the clips containing headshake based on the annotation for headshake described above, using the `split_elan_videos` script (Börstell, 2022) in R version 4.3.2 (R Core Team, 2022) with RStudio version 2023.12.1 (Posit team, 2024). The details can be found in the RMarkdown document following this link: <https://osf.io/mxvre/>.

The clips were then analyzed in OpenFace (Baltrusaitis et al., 2018). OpenFace is a toolkit for face landmark detection, head pose estimation, and facial action unit recognition. Most relevant for this project is that OpenFace measures per frame head rotation along three axes (pitch, roll, and yaw) in radians. Headshake is essentially yaw rotation, labeled as `pose_Ry` in OpenFace. We use the `pose_Ry` measure to measure headshake.

OpenFace also estimates confidence of the measurement (once per frame), which allowed us to filter out the data points with confidence below 0.8. We also excluded four examples of headshake which contain a turn in the middle of the headshake changing the base head position (e.g. because the signer is turning towards a different interlocutor), as these turns would incorrectly affect the phonetic measures. After the clean-up phase, 215 instances of headshake remain in the data set.

### 3.4. Phonetic measurements

Partially following the procedure from Chizhikova and Kimmelman (2022), we decided to explore a wide variety of phonetic measures of the headshakes, which can be divided into two major groups: those requiring peak identification, and those not requiring peak identification.

The measures without peak identification include duration, rough amplitude (the difference between the maximum and the minimum `pose_Ry`), mean velocity (measured as the average difference between two adjacent frames) and peak velocity (measured as the maximum difference between two adjacent frames).

The other measures require identifying the peaks (which reflect the maximally turned positions of the head). As discussed also by Chizhikova and Kimmelman (2022), peak identification algorithms have a sensitivity parameter that needs to be calibrated in order to not identify extremely small local peaks which are due to noise in the OpenFace outputs and do not reflect real changes in head movement direction. Using manual testing and graphical exploration of the data, we determined

the appropriate sensitivity parameter at 0.02 radians. We also decided to include the first and last frames as peaks manually (if not already recognized as such by the algorithm) in order to measure the difference between these and adjacent peaks. Figure 1 illustrates a single headshake with peaks identified. For the details, please see <https://osf.io/mxvre/>.

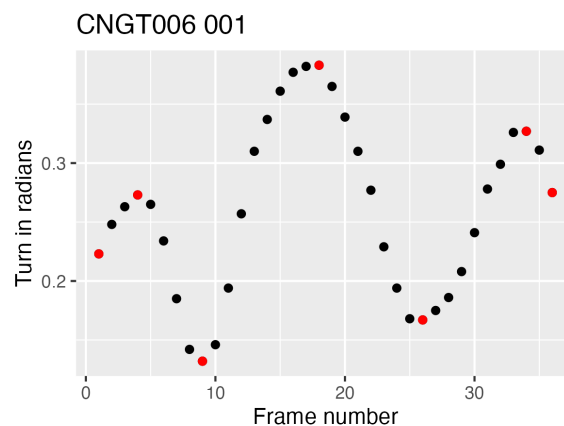


Figure 1: Results of peak identification in one headshake. Red dots are identified peaks.

Once we have the peaks identified, we can derive the following measures: peak number, frequency (number of peaks per second) and amplitude (measured as the average difference between adjacent peaks within a headshake).

An important issue concerns the boundaries of the annotations for headshake types. During the annotation process, we aligned these boundaries with the boundaries of the corresponding manual signs. However, this means that some parts of the headshake overlapping with transitional movements are excluded. Therefore, we also recorded the data such that these parts of the headshake are split equally between the adjacent manual signs. We conducted the analysis described below using both approaches. The general trends are the same between the two approaches, but the effects are less pronounced with the extended annotations.

### 3.5. Statistical analysis

In order to investigate the influence of linguistic functions on the phonetic properties of headshake, we explore numerically and graphically the relation between the three linguistic functions and the phonetic measures, using R version 4.3.2 (R Core Team, 2022) with RStudio version 2023.12.1 (Posit team, 2024). In each case, we calculate the mean and sd estimates per group, create violin and boxplots to explore the relation, and build mixed effect linear regression models, with individual signers coded

as random factors. The full script can be found following this link: <https://osf.io/mxvre/>.

An important disclaimer that we want to make is that the design of the study is inherently exploratory. We try out multiple phonetic measures as we do not have a solid reason to choose on of them beforehand. For example, both rough amplitude and amplitude are measures of amplitude (the size of movement), and mean and peak velocity both measure the speed of movement. There is therefore a higher chance that some of the findings which are reported as significant, are in fact accidental. We interpret the significant differences simply as indication of where effects might be, which need to be further investigated in the future.

## 4. Results

### 4.1. Overall results

Overall, the headshakes in the dataset are characterized by the following measures of central tendencies, reported in Table 1.

measures	mean	median	sd
duration (ms)	25.5	22	13.5
rough amplitude (rad)	0.25	0.21	0.16
mean velocity (rad/sec)	0.03	0.02	0.02
peak velocity (rad/sec)	0.08	0.06	0.06
N peaks	5.93	5.00	3.18
frequency (turns/sec)	6.25	6.06	1.86
amplitude (rad)	0.14	0.11	0.10

Table 1: Central tendencies of the phonetic measures of headshakes.

Thus we can see, for example, that the average duration of a headshake is around 25 frames (1s), the average number of peaks (turns) is almost 6, with an average frequency of 6 turns per second. For all the measures, the mean is higher than the median, so the distributions are positively skewed, with the majority of the data in the lower part of the distribution, and some outliers at the higher end.

Comparing the results with RSL (Chizhikova and Kimmelman, 2022), we can notice that the rough amplitude is comparable between the two languages (0.25 radians in NGT vs. 0.28 in RSL), but that frequency is lower in NGT (6.25 vs. 7.9 Hz). Note, however, that the methodologies used in the two studies are not identical.

### 4.2. Manual negation and spreading

Our NGT dataset includes sentences both with and without manual negative signs, and sentences with and without spreading of the headshake. Both of these factors can potentially influence the phonetic

properties of the headshake.<sup>3</sup>

Not surprisingly, spreading significantly affects the duration of the headshake (headshakes with spreading are longer by an estimated 11 frames<sup>4</sup>), while the presence of a manual negative sign does not affect the duration (Figure 2).

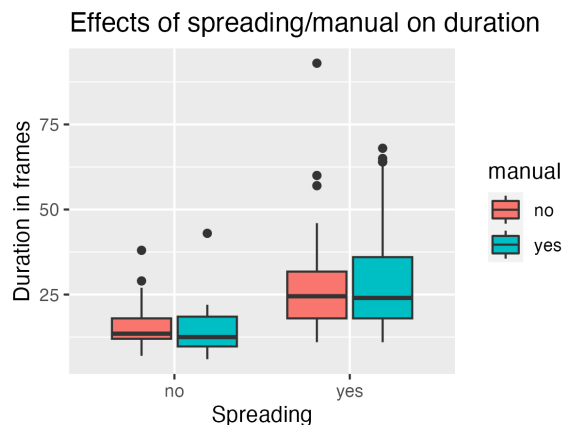


Figure 2: Effect of spreading and the presence of a manual negator on duration.

At the same time, rough amplitude, mean velocity and peak velocity are not affected by spreading or the presence of a manual negative sign.

Turning to the measures based on peak identification, again, not surprisingly, headshakes with spreading have higher number of peaks (turns), by an estimated 2.2 peaks, while the presence of a manual negator does not play a role. More surprisingly, headshakes with spreading have a lower frequency (estimated -1Hz), in comparison to those without spreading. This can be explained as follows: the cases of headshakes without spreading are quite short, but they still need to fit enough turns to be visually salient, and this leads to them having higher frequency. The peak-based amplitude measurement is not significantly affected by spreading or the presence of a manual negator.

### 4.3. Headshake types

Based on the framework discussed above, we divide the headshake into lexical, morphological, and prosodic parts, based on the type of sign it co-occurs with, cf. (3). We expect lexical and possibly morphological headshake to be more phonetically/prosodically prominent as these two types realize syntactic/semantic features; we do not have

<sup>3</sup>The nonmanual nonspreading case means that the headshake only accompanies the verb, that is, it is morphological headshake in our approach.

<sup>4</sup>The estimates here and below are based on the mixed effects model predictions.



a clear prediction on the lexical vs. morphological headshake.<sup>5</sup>

For the non-peak based measures, duration and mean velocity do not significantly correlate with the headshake types. However, both rough amplitude and peak velocity show a difference in the expected direction. When comparing prosodic headshake with the other two types combined, we observe a lower amplitude (by estimated -0.024 radians) and a lower peak velocity (by -0.007 radians per frame). Thus, we find evidence in favor of the hypothesis that prosodic headshake has a weaker realization. Note, however, that the differences, albeit significant, are very small. We do not find a significant difference between lexical and morphological headshake.

Turning to the peak-based measures, the number of peaks is not different for the different categories. However, prosodic headshake has a significantly higher frequency than the other two types (by estimated 2.2 Hz), and a significantly lower peak-based amplitude (by estimated -0.019 radians), which is in agreement with (and equally small as) the result for the rough amplitude measure above. We do not have a clear explanation for the higher frequency of prosodic headshake, but we can hypothesize that since the amplitude is decreased, a higher number of turns can be realized with the same effort in the same time period. We do not find a significant difference between lexical and morphological headshake.

#### 4.4. Negative signs

The type of negative sign might also potentially correlate with phonetic measures of the headshake. Here, however, we do not have a clear prediction beforehand, and thus simply explore the phonetic properties of the four types. Note also that the Neg.Comp type only includes 5 cases, so any conclusions for this type are very tentative.

Of all the measures we applied, only duration and frequency produce significant results, and only for the Neg.Comp type (which is longer and has a lower frequency than standard negation Neg). However, given the extremely small number of data points, we have to conclude that we simply do not have enough data to seriously address this question. For the three types of negative signs with larger number of data points (Neg, Neg.Mod and Neg.Word), we do not see significant differences for any of the measures, resembling the findings from RSL (Chizhikova and Kimmelman, 2022).

<sup>5</sup>Here we report the results obtained using the boundaries aligned with the manual signs, and not the extended boundaries.

#### 4.5. Overall amplitude development

When exploring the effects of the linguistic factors on amplitude, we also noticed a potential general trend of amplitude development over time. This trend is shown in Figure 3, where we plot the average amplitude difference between two adjacent turns ( $\pm 2$  standard errors) for the turn positions. In other words, the figure shows how large the first, second, third, etc. turns are on average.

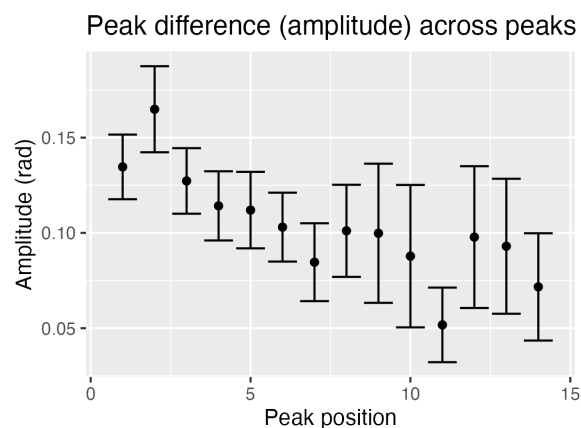


Figure 3: Mean difference in amplitude between adjacent peaks based on peak position. Error bars indicate  $\pm 2$  standard errors.

This figure indicates that the overall trend of amplitude development is as follows: the maximal turn happens at the second position, and then the amplitude goes down steadily (note that we count the neutral position in the beginning of the headshake as the first peak, and thus it is not surprising that the first turn, which is in fact a half turn, is smaller than the second).

It is possible to see a parallel here with down-step or declination of pitch in spoken languages (Pierrehumbert, 1980). Even though amplitude is apparently used for linguistic purposes (distinguishing headshake types), the overall trend is that the highest effort, and thus the highest amplitude, happens in the beginning of the utterance and declines toward the end. However, this issue needs to be studied in much more detail.

## 5. Discussion

### 5.1. Methodological aspects

Similar to Chizhikova and Kimmelman (2022), we show that it is possible to use OpenFace to measure headshake in sign languages, and to investigate the phonetic properties of these headshakes. However, it is important to realize that substantial data processing and semi-manual clean up is required.

First, it is necessary to identify the headshakes during the annotation phase. Second, cases where other non-negative head turns co-occur with the negative turns have to be excluded.

Third, the peak-identification algorithm needs to be manually calibrated. In addition, the same calibration might not work for 100% of the cases. We intend to further explore and improve the peak identification approach in future studies.

Finally, for the analysis of headshake types and other research questions involving overlap with annotations for manual signs, it is not clear how to identify the relevant region, specifically, whether the transitional movements should be included. In our data, it seems that including transitional movements leads to less clear results.

To explore phonetic properties of headshakes, we used a wide variety of measures, some of which are very similar (the two types of amplitude), and some of which are causally related to others (number of peaks, duration and frequency). From our exploration at least, we can conclude that the two measures of amplitude are pretty similar and produce similar results. Given that rough amplitude does not require peak identification, it might be a more practical measure. However, it is not usable for research question involving amplitude dynamics, as discussed in Section 4.5. Mean velocity and peak velocity are also quite similar, but peak velocity seems to be more sensitive (in our data).

## 5.2. Theoretical implications

Keeping in mind the exploratory nature of the study, we can still report some interesting and theoretically consequential findings.

First, we found that, unsurprisingly, spreading headshakes are longer and have a higher number of turns than non-spreading headshakes. More surprisingly, non-spreading headshakes have a higher frequency, which can be a compensatory mechanism in order to make the short non-spreading headshake more saliently visible.

It is also quite interesting that the presence or absence of a manual negative sign does not appear to play a role in any phonetic features of the headshake. This is not fully expected, as the manual sign in some sense renders the headshake superfluous. It might indicate that, in non-manual dominant sign languages like NGT, the non-manual marker is in fact primary, and thus, it is the manual sign that is superfluous and therefore less influential.

The potentially most exciting results concern the headshake types. In agreement with the theory presented in Section 2.2, headshake behaves differently depending on the manual sign it co-occurs with. Prosodic parts of headshake are realized with

smaller amplitude and smaller peak velocity, in comparison to the morphological and lexical parts. This is a clear demonstration that syntactic factors affect the realization of the negative headshake, and, to our knowledge, the first demonstration of this type of effect for headshake in SLs. Note however, that the differences in amplitude and velocity are very small relative to the overall mean amplitude and peak velocity.

Given the methodological challenges and the exploratory nature of this study, we cannot be fully confident in our findings, but we think that the study provides a good indication that future studies on phonetic properties of non-manual markers using Computer Vision can be expected to lead to interesting discoveries.

## Author Contributions

**Vadim Kimmelman:** Conceptualization, Funding Acquisition, Methodology, Investigation, Formal Analysis, Visualization, Software, Writing. **Marloes Oomen:** Conceptualization, Methodology, Investigation, Writing. **Roland Pfau:** Conceptualization, Methodology, Investigation, Writing.

## Acknowledgements

This project is funded by the European Union (ERC, Nonmanual, project number 101039378, awarded to V. Kimmelman). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. Oomen's contribution to this project is funded by the Dutch Research Council (NWO, project number VI.Veni.211C.052, awarded to M. Oomen).

## 6. Bibliographical References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.
- Carl Börstell. 2022. R functions for working with linguistic data (mapping, ELAN, video processing, etc.).
- Michele Brunelli. 2011. *Antisymmetry and Sign Languages: A Comparison between NGT and LIS*. Ph.D. thesis, University of Amsterdam, Amsterdam.

- Svetlana Burkova. 2015. The ways of expressing nominal plurality in the Russian sign language. *Sibirskij fonologičeskij zhurnal*, 2:174–184.
- Anastasia Chizhikova and Vadim Kimmelman. 2022. Phonetics of Negative Headshake in Russian Sign Language: A Small-Scale Corpus Study. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 29–36, Marseille, France. European Language Resources Association (ELRA).
- Jane Coerts. 1992. *Nonmanual Grammatical Markers. An Analysis of Interrogatives, Negation and Topicalisation in Sign Language of the Netherlands*. Ph.D. thesis, University of Amsterdam, Amsterdam.
- Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43. ELRA, Paris.
- Onno Crasborn, Inge Zwitterlood, and Johan Ros. 2008. Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands. <http://www.ru.nl/corpusngtuk/introduction/welcome/>.
- Kadir Gökgöz. 2021. Negation: Theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 266–294. Routledge, London; New York.
- Jessica Harmon. 2017. Simultaneous articulation as a window into structure: Nonmanuals in ASL. In Julia Nee, Margaret Cychosz, Dmetri Hayes, Tyler Lau, and Emily Ramirez, editors, *Proceedings of the Forth-Third Annual Meeting of the Berkeley Linguistics Society*, volume I, pages 121–144. Berkeley Linguistics Society, Berkeley.
- Simon Harrison. 2014. [The organisation of kinesic ensembles associated with negation](#). *Gesture*, 14(2):117–140.
- Trevor Johnston. 2018. [A corpus-based study of the role of headshaking in negation in auslan \(australian sign language\): Implications for signed language typology](#). *Linguistic Typology*, 22(2):185–231.
- Adam Kendon. 2002. [Some uses of the head shake](#). *Gesture*, 2(2):147–182.
- Ulrika Klomp, Marloes Oomen, and Roland Pfau. in press. Negation in Sign Language of the Netherlands. In Veselinova and M. Miestamo, editors, *Negation in the Languages of the World*. Language Science Press, Berlin.
- Hannah Lutzenberger, Roland Pfau, and Connie de Vos. 2022. [Emergence or Grammaticalization? The Case of Negation in Kata Kolok](#). *Languages*, 7(1):23.
- Bahtiyar Makaroğlu. 2021. [A Corpus-Based Typology of Negation Strategies in Turkish Sign Language](#). *Dilbilim Araştırmaları Dergisi*, 32(2):111–147.
- Marloes Oomen and Roland Pfau. 2017. [Signing not \(or not\): A typological perspective on standard negation in Sign Language of the Netherlands](#). *Linguistic Typology*, 21(1):1–51.
- Marloes Oomen, Roland Pfau, and Enoch Aboh. 2018. [High and low negation in Sign Language of the Netherlands](#). *FEAST. Formal and Experimental Advances in Sign language Theory*, (1):39–47.
- Roland Pfau. 2015. [The grammaticalization of headshakes: From head movement to negative head](#). In Andrew D.M. Smith, Graeme Trousdale, and Richard WALTEREIT, editors, *New Directions in Grammaticalization Research*, pages 9–50. John Benjamins Publishing Company.
- Roland Pfau. 2016. A Featural Approach to Sign Language Negation. In Pierre Larrivé and Chungmin Lee, editors, *Negation and Polarity: Experimental Perspectives*, pages 45–74. Springer International Publishing, Cham.
- Roland Pfau, Tamar Makharoblidze, and Hedde Zeijlstra. 2022. [Negation and Negative Concord in Georgian Sign Language](#). *Frontiers in Psychology*, 13:734845.
- Roland Pfau and Josep Quer. 2002. V-to-Neg raising and negative concord in three sign languages. *Rivista di Grammatica Generativa*, 27:73–86.
- Roland Pfau and Josep Quer. 2010. Nonmanuals: Their prosodic and grammatical roles. In Diane Brentari, editor, *Sign Languages*, pages 381–402. Cambridge University Press, Cambridge.
- Janet Breckenridge Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology.
- Posit team. 2024. *RStudio: Integrated Development Environment for R*. Boston, MA.

- Josep Quer. 2012. Negation. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign Language: An International Handbook*, pages 316–339. De Gruyter Mouton, Berlin.
- R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Pavel Rudnev and Anna Kuznetsova. 2021. [Linearization constraints on sentential negation in Russian Sign Language are prosodic](#). *Sign Language & Linguistics*, 24(2):259–273.
- Joke Schuit. 2013. *Typological Aspects of Inuit Sign Language*. Ph.D. thesis, University of Amsterdam, Amsterdam.
- Cindy Van Boven, Marloes Oomen, Roland Pfau, and Lotte Rusch. 2023. [Negative Concord in Sign Language of the Netherlands: A journey through a corpus](#). In Ella Wehrmeyer, editor, *Studies in Corpus Linguistics*, volume 108, pages 30–65. John Benjamins Publishing Company, Amsterdam.
- Ronnie Wilbur. 2021. Non-manual markers: Theoretical and experimental perspectives. In Josep Quer, Roland Pfau, and Annika Herrmann, editors, *The Routledge Handbook of Theoretical and Experimental Sign Language Research*, pages 530–565. Routledge, London; New York.
- Ronnie B. Wilbur and Cynthia Patschke. 1999. [Syntactic Correlates of Brow Raise in ASL](#). *Sign Language & Linguistics*, 2(1):3–41.
- Jun Hui Yang and Susan Fischer. 2002. Expressing negation in Chinese Sign Language. *Sign Language & Linguistics*, 5(2):167–202.
- Ulrike Zeshan. 2004. [Hand, head, and face: Negative constructions in sign languages](#). *Linguistic Typology*, 8(1):1–58.
- Ulrike Zeshan, editor. 2006. *Interrogative and Negative Constructions in Sign Languages*. Number 1 in Sign Language Typology Series. Ishara Press, Nijmegen.



# Nonmanual Marking of Questions in Balinese Homesign Interactions: a Computer-Vision Assisted Analysis

Vadim Kimmelman<sup>1</sup> , Ari Price<sup>2</sup> , Josefina Safar<sup>3</sup> ,  
Connie de Vos<sup>3</sup> , Jan Bulla<sup>1,4</sup> 

<sup>1</sup>University of Bergen, <sup>2</sup>Western Norway University of Applied Sciences,  
<sup>3</sup>Tilburg University, <sup>4</sup>University of Regensburg  
<sup>1,2</sup>Bergen, Norway, <sup>3</sup>Tilburg, the Netherlands, <sup>4</sup>Regensburg, Germany,  
{vadim.kimmelman, jan.bulla}@uib.no, julie.anna.price@hvl.no,  
{C.L.G.deVos, J.Safar}@tilburguniversity.edu

## Abstract

In recent years, both linguistic resources and computer-based tools have been developed that make it possible to investigate research questions that have not been studied before. In this study, we conduct a study of nonmanual question marking, using data from the Balinese Homesign Corpus – a unique resource documenting language use in several Balinese homesigners. We further demonstrate how using OpenFace, a Computer-Vision solution, allows for quantitative analysis of head tilts used by these signers in marking questions. We also showcase a pilot statistical analysis of the dynamic kinetic contours of the head movements.

**Keywords:** Balinese Homesign Corpus, questions, nonmanual marking, head tilt, OpenFace

## 1. Introduction

In recent years, both linguistic resources and computer-based tools have been developed that make it possible to investigate research questions that have not been studied before. In this study, we illustrate such a linguistic resource, namely the Balinese Homesign Corpus, containing unique data of several homesigners in conversations with their family members and other signers. This data set allows investigating nonmanual marking of questions in this special population, that is, deaf individuals who grew up without access to an already existing signed language. As we will discuss below, the homesigners are not in contact with each other, so their homesign systems are potentially completely distinct. At the same time, they co-create their homesign systems with their hearing family members, who are representatives of the local hearing community. Therefore, it is possible that the homesign systems will partially converge, also in the domain of question marking, due to the influence of the gestures of the hearing non-signers.

In addition, in this paper we explore the use of one of the relatively novel tools from the Computer Vision domain, namely OpenFace (Baltrušaitis et al., 2018). This tool allows identification and tracking of the head and the facial features, including measurements of the rotation of the head. Since, as we show, head pitch (up and down movement) are important markers of questions in our data set, we use the tool to measure pitch and correlate it with specific question types across the different signers in the corpus.

Finally, using a subset of the data, we show

how OpenFace measurements can potentially be used to study dynamic kinematic properties of head movements in more detail, using various smoothing techniques. While this type of analysis will require extensive follow-up research, we demonstrate the promise that it has.

### 1.1. Question Marking in Sign Languages

Question marking has been studied for many sign languages (Zeshan, 2004; Cecchetto, 2012). Almost universally, nonmanual markers are employed in marking questions of different types, especially for polar questions (ibid.). For content questions, in many sign languages question words are used, often also accompanied by nonmanual markers. Quite strikingly, in a majority of sign languages, polar questions are accompanied by raised eyebrows, while content questions have more diverse patterns of marking. In addition, many studies report different types of head tilts marking for questions (Cecchetto, 2012). In her thesis about an emerging sign language in Brazil, Fusellier-Souza (2004, 304) specifically categorizes eyebrow raises and head tilts as modality-specific traits that function as non-manual features of question marking, in addition to expressing doubts and uncertainty, across sign languages. Very little is known about question marking in homesigners, with the exception of a case study on David, a child homesigner from the US, who has been reported to only use a manual flip gesture to mark wh-questions (Franklin et al., 2011). Based on the existing research, we thus decided to focus specifically on potential nonmanual markers in the homesign data described below.

## 1.2. Homesigners

Homesign is a visual-gestural communication system that is co-created by a deaf person who does not have full access to a conventionalized language and the attentive interlocutors in their proximity (De Vos, 2023). Observing homesign systems allows for an opportunity to view aspects of emerging linguistic systems that can provide insights into human language development. Each homesign system can have unique features, as with more conventionalized languages, but there are some elements that have been considered resilient properties of language (Brentari and Goldin-Meadow, 2017). What classifies these features as resilient is that they show up across different home sign systems, which all lack a distinct input from a pre-existing language model. However, the form in which the functional feature may present itself can vary from one homesign system to another. This paper seeks to explore the forms that different homesigners, who are not in contact with each other but come from the same cultural background, use to mark question types with nonmanual markers typically found across sign languages. In particular, we observe the use of upward and downward head tilts across question types in conversational data between homesigners and their interlocutors.

## 1.3. Computer Vision Analysis of Nonmanuals

An important goal of this study is to test the applicability of Computer Vision tools to linguistic analysis of nonmanuals. In recent years, due to success of the Deep Learning approach to Computer Vision, several toolkits for detecting and tracking body landmarks in video recording have appeared, including OpenPose (Cao et al., 2018) and MediaPipe (Lugaresi et al., 2019). Some of the tools also include automatic 3D reconstruction from 2D video recordings, and tracking head rotation, such as OpenFace (Baltrusaitis et al., 2018), which we use here for this reason.

First studies using Computer Vision tools to analyze sign languages have appeared over 10 years ago (Metaxas et al., 2012; Karppa et al., 2014). However, due to the relative user-friendliness of the new tools and their increased reliability and efficiency, in recent years, a large number of publications applying them to sign language and gesture data has appeared (see for example Östling et al. 2018; Trujillo et al. 2019; Fragkiadakis 2022; Börstell 2023), also for analyzing nonmanual markers (Kimmelman et al., 2020; Chizhikova and Kimmelman, 2022). While the use of these tools for sign language analysis is very promising, extensive testing and calibration of these tools is required (Kuznetsova et al., 2021); at this stage, it is neces-

sary to combine these tools with manual annotations, as we do in the current study.

## 2. Methodology

### 2.1. The data

This data set consists of 5 videos from the [Balinese Homesign Corpus](#). The videos contained recordings of 11 people who had experience using a homesign system. The participants were 5 prelingually deaf homesigners, who do not have input from an adult sign language model, 5 hearing interlocutors and 1 deaf interlocutor with knowledge of a conventionalized sign language (Indonesian Sign Language (BISINDO)). All homesigners and their interlocutors in our data set come from the Buleleng regency in Northern Bali, Indonesia. Due to their regional proximity, the homesigners and their interlocutors have a similar cultural background, which includes shared knowledge of locations, rituals, and traditional family systems among other norms. Having a common culture also gives these homesigners access to the gestural repertoire affiliated with the larger local community of hearing speakers of Balinese. The data was collected by a team of hearing and deaf research assistants in Bali. Each conversation was filmed with two Canon HF G50 cameras and conversations lasted from 10-50 minutes. This resulted in a total of 02:24:45 worth of video footage.



Figure 1: Homesigner HS01 (right) asking a polar question, in conversation with her mother.

Most of the deaf homesigners were filmed having a conversation with a hearing relative, such as their mother, sister, or sister-in-law, with one having two hearing relatives present (see Figure 1 for an illustration). One deaf homesigner was filmed signing with another deaf person, a man from a neighbouring village who he was not related to, but had met several times previously. This deaf man

attended deaf school for a number of years, where he acquired BISINDO. Thus, all deaf homesigners without long-term formal schooling interacted with an interlocutor who knew a conventionalized language in addition to using homesign. However, participants differed in terms of their ages (27-53), professions, and marital statuses, which influenced conversation topics. For more detailed information, please see [Safar and De Vos \(2022\)](#), who used the same data set.

## 2.2. Annotation

The videos were annotated in ELAN 6.7 ([Crasborn and Sloetjes, 2008](#)). Previous annotations done by [Safar and De Vos \(2022\)](#) provided an English translation tier that acted as a baseline to pinpoint questions in the video data set. Expanding upon the original files from [Safar and De Vos \(2022\)](#), a tier was added to mark where questions came up in the interactions of homesigners. On this tier, question types were then marked as being 1 of 4 types: 'polar,' 'open,' 'content,' or 'huh'. While 'huh' is not necessarily a question in and of itself, it proved to have a fairly consistent form across homesigners and provided a similar function to an open question by prompting the interlocutor to give more information. The category of 'content' question was used when a manual sign (question word) was used, while 'open' question do not contain a manual question word, but instead a gap in place of one of the constituents, and presuppose the answer to fill in this gap.

After marking the question types, the 'NMM-annotation-template.etf' template created by [Oomen et al. \(2023\)](#) was imported into the original ELAN files to allow for the consistent annotation of nonmanual markers across homesigners. Following [Oomen et al. \(2023\)](#), the nonmanual markers in each question were annotated, with special attention given to head position and eyebrows. In particular, up and down head movements were marked on the 'NMM.head-y' tier and more rapid movements were marked as nods on the 'NMM.head-move' tier. Raised, neutral and lowered eyebrow movements were also then marked on the 'NMM.eyebrows' tier. In order to explore the nonmanual markers of these signers as individuals, separate files were made for each signer in the data set, except for a 'third participant' in one video that did not actively participate in the conversation.

All the new annotations for this study were created by one of the authors, AP. Another author, VK, reviewed the annotations, and AP and VK discussed all the instances of disagreement.

## 2.3. Computer Vision Processing

We extracted the clips containing up and down head movements based on the annotation on the 'NMM.head-y' tier, using the `split_elan_videos` script ([Börstell, 2022](#)). Because the video recording contained two or three signers simultaneously, we cropped the clips to have only one signer in one clip with `ffmpeg`, ([Tomar, 2006](#)). The details can be found in the RMarkdown document in the repository linked below.

The clips were then analyzed in OpenFace ([Baltrusaitis et al., 2018](#)). OpenFace is a toolkit for face landmark detection, head pose estimation, and facial action unit recognition. Most relevant for this project is that OpenFace measures per frame head rotation along three axes (pitch, roll, and yaw) in radians. Up and down head movement is essentially pitch rotation, labeled as `pose_Rx` in OpenFace. We use the `pose_Rx` to measure head movements.<sup>1</sup> OpenFace also estimates confidence of the measurement (once per frame), and so we filter out the data points with confidence below 0.9.

## 2.4. Statistical Analysis

The full documentation of the statistical analysis and the data files used for the analysis can be found in this repository: <https://osf.io/5d7wu/>.

### 2.4.1. Analyzing the Annotations

As the first step, we graphically explore the relations between question type and the nonmanual markers (eyebrow movements, head movements, head pitch), both overall and for individual signers. We also compare the deaf signers to their hearing interlocutors. The analysis was conducted in R ([R Core Team, 2022](#)) with RStudio ([Posit team, 2024](#)), using `tidyverse` ([Wickham et al., 2019](#)) and `ggplot2` ([Wickham and Chang, 2016](#)).

### 2.4.2. Analyzing OpenFace Outputs

We used OpenFace to extract measurements of head pitch (`pose_Rx`). First, we investigated the relation between our annotations for head movement and the measures outputted by OpenFace in order to see whether they generally agree. After establishing that this is indeed the case, we investigated the relation between the OpenFace pitch

---

<sup>1</sup>OpenFace also tracks the eyebrows and even automatically detects eyebrow raise. However, as previous research has shown, these measures are very unreliable in the presence of head tilts ([Kuznetsova et al., 2021](#)). We have also tested them with this data set and came to the same conclusion: eyebrow measures from OpenFace cannot be used for linguistic analysis, at least not for question marking.



measurements and our question type annotations. We used the same tools as above, with addition of the `lme4` package (Bates et al., 2015) for mixed effect regression.

### 2.4.3. Analyzing Specific Dynamic Movements

The head movements (both up and down) are dynamic movements, and our long-term goal is to investigate them as such, and not as average measures per movement as in the previous section. In order to start developing this approach, we have selected 24 up and down movements produced by HS01 and investigated them further.

The observations collected by the variable `pose_Rx` represent a continuous movement, but are of discrete nature. In addition, one has to assume that the recorded movements contain a certain amount of noise from the recording process. Last and maybe most importantly, head movements do not always follow a precise identical patterns, but may overlap with smaller movements which are – in principle – negligible.

We therefore process the head movement observations with three different non-parametric statistical methods that permit to detect general patterns in noisy data: locally estimated scatterplot smoothing (LOESS), kernel regression, and splines using the statistical software package R (R Core Team, 2021), version 4.05.

LOESS smoothing (Shyu and Cleveland, 1992) is based on local polynomial regression. It is available through the function `loess`, which uses polynomials of degree two by default. Moreover, LOESS requires input of the span parameter, which controls the degree of smoothing. We determined this parameter by 10-fold cross-validation with mean absolute error as criterion.

The core of Kernel regression is the Nadaraya–Watson estimate estimator (Watson, 1964; Nadaraya, 1964), available in R within the `np` package (Helwig, 2021). This estimator relies on input of an optimal bandwidth parameter, which determines the degree of smoothing. We chose Kullback-Leibler cross-validation (Hurvich et al., 1998) in the `npregbw` function for this task, because the default least-squares cross-validation turned out to be too wiggly.

A wide range of implementations exists for spline regression. All have in common that the shape of the resulting function mainly depends on the number (and placement) of knots, and a smoothing parameter. We considered i) the function `ss` from the `np` package with the default generalized cross-validation for choosing the smoothing parameter; ii) the `gam` function with default settings from the `gam` package (Hastie and Tibshirani, 1990; Chambers

and Hastie, 1992); iii) p-splines via the function `gam` from the `mgcv` package.

## 3. Results

### 3.1. General findings

In total, we annotated 296 examples of questions in the data. However, the data is very unbalanced. First, 215 (73%) of the questions are polar questions. Second, different individuals produced drastically different numbers of examples. In fact, all but one examples of the *huh?* type were produced by a single hearing participant, and most examples of open questions were produced by another hearing participant. It is therefore difficult to make any generalization based on this data. However, it is important to remember that the individuals in the data set do not represent a population of users of a single language. Instead, each homesigner (possibly with their hearing family members) represents a completely unique system. If we discover some general tendency despite the skewed data sample and despite the potential differences between the homesign systems, it is even more surprising.

### 3.2. Annotation-Based Analysis

We explore the patterns of nonmanual marking by plotting the annotations for eyebrow movement, head movement and head pitch in relation to the type of question they overlap with.

In Figure 2 we can see the eyebrow movement patterns across the question types and the different signers.<sup>2</sup> It is clear that there is great variation between the signers. Focusing on the signers with the most data, HS01 deaf signer raised her eyebrows consistently for both polar and content questions; HS10's hearing conversational partner raised the eyebrows in open and content questions, but less frequently so for polar questions, and the HS17's deaf conversational partner (Deafb on the Figure) basically did not raise his eyebrows at all. One general pattern that emerges from these signers is that eyebrow marking is more varied for the polar questions than for the other types.

In Figure 3 we can see that again, there is a lot of variation between the signers, but something that is noticeable is that almost all the signers use head nods for polar questions, and less so for the other

<sup>2</sup>On this and the following Figures, the codes for individual signers consist of two parts. The first part refers to the conversation code in the corpus; the second part specifies whether the signer was deaf or hearing, with additional letters distinguishing the two deaf signers in one of the conversations.



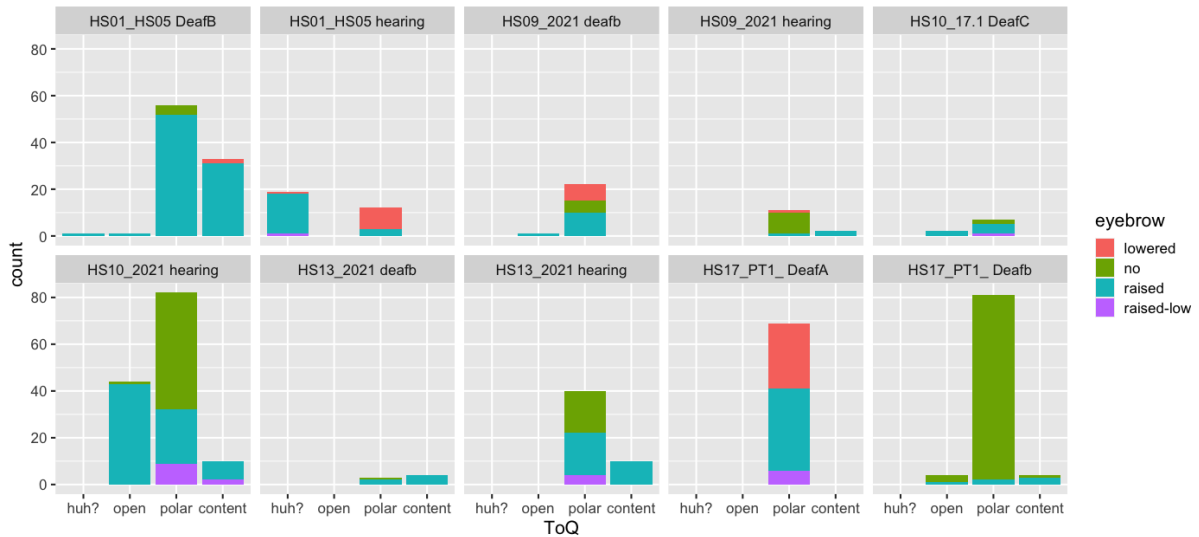


Figure 2: Eyebrow movements across question types, for each individual signer.



Figure 3: Head movements across question types, for each individual signer. Most relevant colors: brown: nods, pink: no movement, purple: headshake.

types.<sup>3</sup> One deaf signer (HS17) uses some headshakes in polar questions, which is explained by the fact that these are questions containing negation.

Finally, in Figure 4, despite the variation, we can see an interesting pattern emerging: the up movement is almost never used for polar questions (which use a lot of pitch down, or no pitch), but very dominantly for the other types. This is even more clear when the data is aggregated for all the signers in Figure 5.

In addition, we have explored whether there is a relation between the markers used by the deaf vs. hearing signers. Overall, we do not find clear differences. One noticeable difference is that, propor-

tionally, the deaf signers had more neutral brow positions in polar questions and produced less down movements, but this is mostly driven by a single signer, as can be seen in Figure 4.

### 3.3. Computer-Vision-Based Analysis

After extracting the pose\_Rx (pitch) measurements with OpenFace, as the first step we analyzed the relation between our annotated categories for pitch (up vs. down vs. neutral labels). The results are visualized in Figure 6.

Thus, as expected, the cases which we annotated as head pitched up have a higher average measurement of pitch in OpenFace than the cases annotated as pitch down, and the neutral cases are

<sup>3</sup>For the clarification of the other less frequent labels, see Oomen et al. (2023)

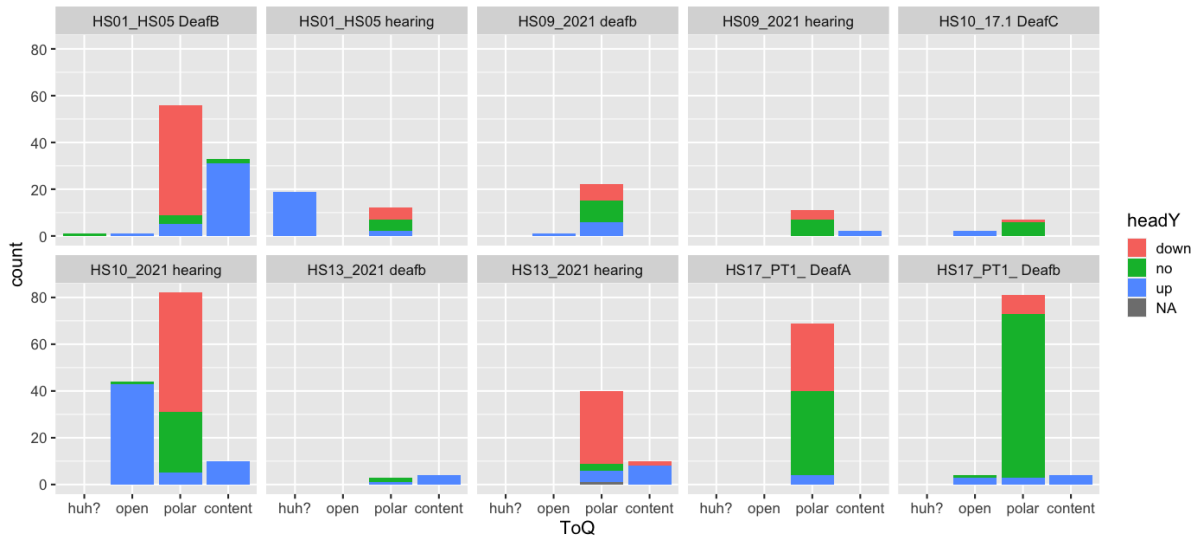


Figure 4: Head pitch across question types, for each individual signer.

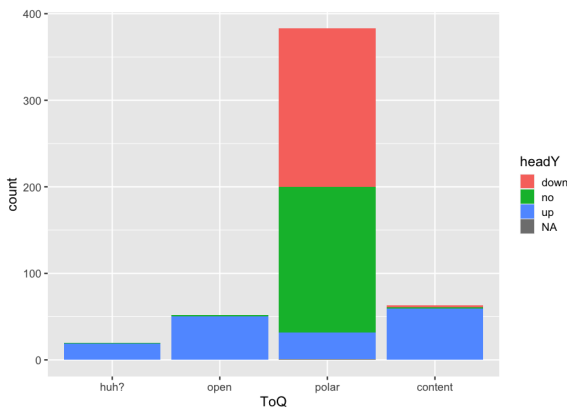


Figure 5: Head pitch across question types, aggregated.

in the middle.

For a more insightful analysis, we also visualize the relation between our annotations for question types, and the overlapping measurements of pitch\_Rx from OpenFace. This is represented in Figure 7.

It is clear that what we find based on our manual annotation is also very visible based on the OpenFace measurements: polar questions on average have a much lower head pitch (the head is moved down), while the other types have a higher pitch. The differences between polar vs. open and polar vs. content are highly significant (polar vs. open estimated difference 0.4 rad,  $p < 0.001$ , polar vs. content estimated difference 0.32 rad,  $p < 0.001$ ), while the difference between polar and huh? is not significant (most likely because almost all instances of huh? are produced by a single signer).

Importantly, the same pattern is visible for the individual signers, modulo the fact that not all of

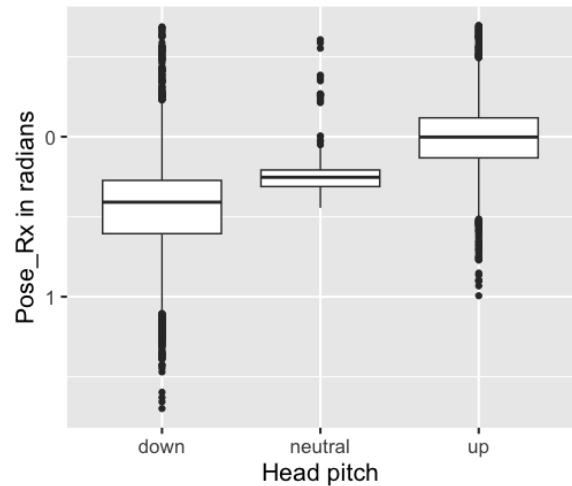


Figure 6: Relation between pose\_Rx and manual annotations for head movement (pitch), aggregated over all the signers. Points beyond the  $\pm 2SD$  removed for visualization purposes.

them have all the question types present.

Thus, the measurements of head pitch from OpenFace produce results agreeing with our observations: polar questions are consistently marked by head down, while the other types of questions are marked with the opposite head movement.

### 3.4. Head Movements as Dynamic Patterns

We selected four typical head movements of varying duration to illustrate the performance of the three smoothing approaches described in Section 2.4.3. Figure 8 illustrates these four movements: the two top panels show a simple upward

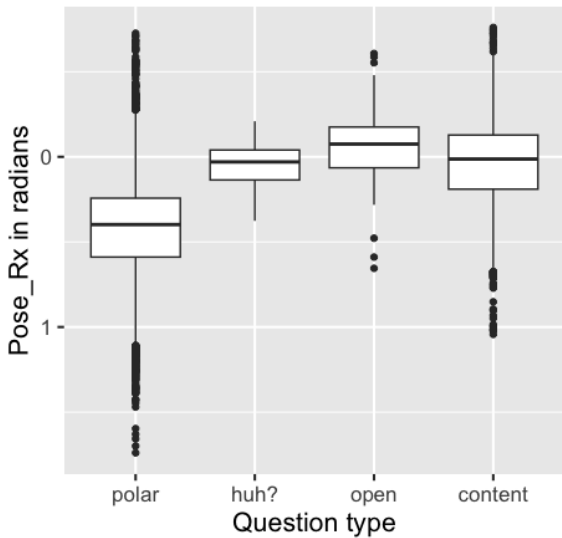


Figure 7: Relation between pose\_Rx and manual annotations for type of question, aggregated over all the signers. Points beyond the  $\pm 2SD$  removed for visualization purposes.

and downward nod, respectively. The two bottom panels present two slightly more complex movements consisting of a double and a multiple nods, respectively. For the simple movements, no large differences are visible between the smoothing methods. Splines obtained from the `gam` package exhibit the highest degrees of smoothing, while those resulting from the `npreg` adapt very (too) closely to the observations. Between these two cases lie the remaining methods, which visually do not differ substantially from each other. The more complex cases in the lower panel paint a more distinct picture: the splines from the `gam` and `mgcv` package do not capture the extent of the movement dynamics sufficiently, in particular for multiple tilt example. This example also illustrates a slight over-smoothing of the LOESS method. However, kernel regression and splines from the `npreg` package reproduce the movement dynamics well, where the latter again provides the highest fit to the observations.

#### 4. Discussion

In this study, we aimed to investigate nonmanual marking of questions produced by Balinese homesigners and their family members in free conversation. We had three main goals: to provide a first description of such marking with attention to variation and similarities between the different signers, to test and explore using OpenFace as a tool to measure head movements in this type of data, and to start exploring analyzing head movements as dynamic patterns using these measurements.

Concerning the first goal, we found a rather inter-

esting pattern. The signers, while not representing a single language, show some degree of convergence on the nonmanual strategies in marking different question types. The eyebrow movements show the most diversity between the signers. This is surprising given the prevalence of eyebrow raise used for polar question marking across different sign languages (Zeshan, 2004). However, the head movements, especially analyzed in terms of head pitch direction (up vs. down) show a surprisingly strong pattern which is similar between all the signers. Specifically, all the signers (both deaf and hearing) mark polar questions with downward pitch, while the other types are more characterized by upward pitch.

The most natural explanation that can be offered for this pattern is that head pitch is used for question marking in similar ways by the surrounding hearing community. This would naturally lead to the hearing family members using these nonmanuals also when signing with their deaf homesigner relatives. It is possible to hypothesize that, for the deaf homesigners and their relatives, the nonmanuals might undergo regularization and become partially obligatory due to their importance in communication. Note that it is clear that head tilt is not universally in other hearing communities, see for example Sze (2022) comparing head tilts in Cantonese speakers with Hong Kong Sign Language signers. So, further research on the nonmanual marking of questions among the surrounding hearing community is required to test this hypothesis.

As for the second goal, it turns out that using OpenFace for analyzing head pitch works very well, at least when averaging the pitch for individual instances of nods/tilts. The measurements agree with our pitch annotations, and there is a strong relation between our annotation for question type and the pitch measurements. Thus, we see an agreement between the manual annotation method and the Computer-Vision based method. Neither method can be considered fully reliable or the ground truth, but it can be a useful methodological improvement to compare and complement the two methods.

When inferring movement dynamics, we observed quite different degrees of smoothing by the various considered methods (see Figure 8). Hereby it should be noted that we mainly relied on default settings of the respective R packages, particularly for the spline regressions. The results suggest that the default number of knots is set too low in both the `gam` and the `mgcv` package, and potentially too high in the `npreg` package. In a couple of additional experiments (not shown) we investigated the effects of modifying various settings in the different packages. It turned out that the spline type has very little effect in our examples, whereas – as to

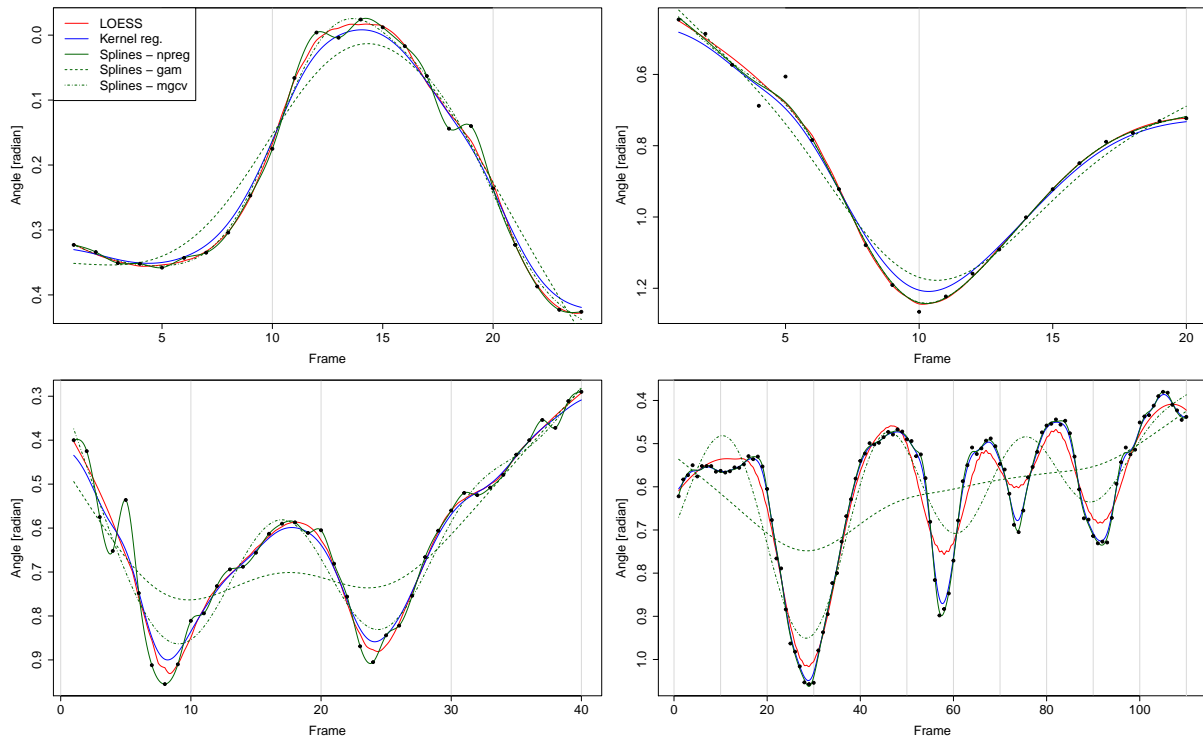


Figure 8: Typical head tilts / movements with inferred dynamics. The  $x$ - and  $y$ -axis show the frame and head tilt angle, respectively. The  $y$ -axis is mirrored for better interpretability. Black dots correspond to the observations, and lines result from fitted models (red: Loess, blue: kernel regression, green: spline regressions).

be expected – the knot number strongly affects the degree of smoothing. Hence, it remains to be investigated whether the performance of the spline regressions can be improved by e.g. optimizing the number of knots through cross-validation or model selection techniques. Furthermore, Kernel regression seems to satisfactorily capture the dynamics in all examples. Last, the cross-validation criteria of LOESS may also be improved, which could help to better describe the most dynamic movements.

Aside from these rather technical aspects, it also remains to investigate how the inferred movement dynamics should be post-processed. Analysis of the inferred curves from Figure 8 should be relatively straightforward by measures such as number of extreme points, duration of movements, or distances between extreme points, to name only a few. Challenges appear, however, from less clear sequences of movements such as displayed in Figure 9. This downward nod between approximately Frame 15 and Frame 35 constitutes the main dynamics of these observations. Problematic are several additional extreme points (Frames 13, 37, 47), which complicate the processing of such a sequence. Either these extremes are not real movements, or they are, but they should not be classified as nods in the linguistic sense. The development and testing of suitable methods constitute topics of

ongoing research.

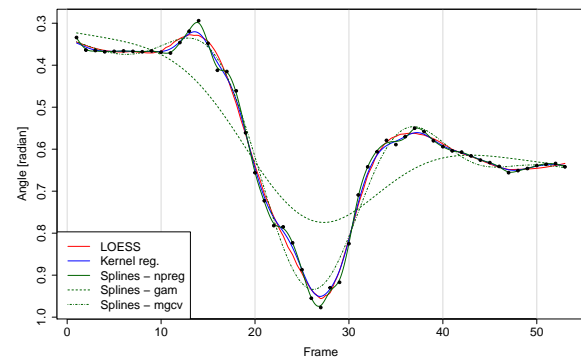


Figure 9: A more complex head tilt / movement with inferred dynamics. The  $x$ - and  $y$ -axis show the frame and head tilt angle, respectively. The  $y$ -axis is mirrored for better interpretability. Black dots correspond to the observations, and lines result from fitted models (red: Loess, blue: kernel regression, green: spline regressions).

## Author Contributions

**Vadim Kimmelman:** Conceptualization, Funding Acquisition, Methodology, Investigation, Formal



Analysis, Visualization, Software, Writing. **Ari Price**: Methodology, Investigation, Writing. **Josefina Safar**: Methodology, Data Curation, Writing. **Connie De Vos**: Methodology, Data Curation, Writing. **Jan Bulla**: Formal Analysis, Visualization, Writing.

## Acknowledgements

This project is funded by the European Union (ERC, Nonmanual, project number 101039378, awarded to V. Kimmelman). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 5. Bibliographical References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE.
- Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. *Fitting Linear Mixed-Effects Models Using lme4*. *Journal of Statistical Software*, 67(1):1–48.
- Diane Brentari and Susan Goldin-Meadow. 2017. *Language Emergence*. *Annual Review of Linguistics*, 3(1):363–388.
- Carl Börstell. 2022. *R functions for working with linguistic data (mapping, ELAN, video processing, etc.)*.
- Carl Börstell. 2023. *Extracting Sign Language Articulation from Videos with MediaPipe*. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 169–178, Tórshavn, Faroe Islands. University of Tartu Library.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.
- Carlo Cecchetto. 2012. Sentence types. In Roland Pfau, Markus Steinbach, and Bencie Woll, editors, *Sign language: An international handbook*, pages 292–315. De Gruyter Mouton, Berlin.
- John M. Chambers and Trevor Hastie. 1992. *Statistical Models in S*. Wadsworth & Brooks/Cole Advanced Books & Software. Google-Books-ID: eKOaQAACAAJ.
- Anastasia Chizhikova and Vadim Kimmelman. 2022. *Phonetics of Negative Headshake in Russian Sign Language: A Small-Scale Corpus Study*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 29–36, Marseille, France. European Language Resources Association (ELRA).
- Onno Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 39–43. ELRA, Paris.
- Connie De Vos. 2023. *Cognitive pragmatics: Insights from homesign conversations*. *Behavioral and Brain Sciences*, 46:e8.
- Manolis Fragkiadakis. 2022. *Assessing an Automated Tool to Quantify Variation in Movement and Location: A Case Study of American Sign Language and Ghanaian Sign Language*. *Sign Language Studies*, 23(1):98–126.
- Amy Franklin, Anastasia Giannakidou, and Susan Goldin-Meadow. 2011. *Negation, questions, and structure building in a homesign system*. *Cognition*, 118(3):398–416.
- Ivani Fusellier-Souza. 2004. *Sémiogenèse des langues des signes. Etude de Langues de Signes Emergentes (LSE) pratiquées par des sourds brésiliens*. Theses, Université Paris 8 - École Doctorale "Cognition, Langage, Interaction" (ED 224). Tex.hal\_id: tel-01701214 tex.hal\_version: v1.
- T. J. Hastie and R. J. Tibshirani. 1990. *Generalized Additive Models*. CRC Press. Google-Books-ID: qa29r1Ze1coC.
- Nate Helwig. 2021. *npreg: Nonparametric Regression via Smoothing Splines*.
- Clifford M. Hurvich, Jeffrey S. Simonoff, and Chih-Ling Tsai. 1998. *Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion*. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(2):271–293.
- Matti Karppa, Ville Viitaniemi, Marcos Luzardo, Jorma Laaksonen, and Tommi Jantunen. 2014. *SLMotion - an extensible sign language oriented*

- video analysis tool. In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1886–1891, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. [Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study](#). *PLOS ONE*, 15(6).
- Anna Kuznetsova, Alfarabi Imashev, Medet Mukushev, Anara Sandygulova, and Vadim Kimmelman. 2021. [Using Computer Vision to Analyze Non-manual Marking of Questions in KRSL](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 49–59, Virtual. Association for Machine Translation in the Americas.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A Framework for Building Perception Pipelines](#).
- Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael, and Carol Neidle. 2012. [Recognition of nonmanual markers in American Sign Language \(ASL\) using non-parametric adaptive 2D-3D face tracking](#). In *8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2414–2420, Istanbul, Turkey. European Language Resources Association (ELRA).
- E. A. Nadaraya. 1964. [On Estimating Regression](#). *Theory of Probability & Its Applications*, 9(1):141–142. Publisher: Society for Industrial and Applied Mathematics.
- M. Oomen, Tobias de Ronde, Lyke Esselink, and Floris Roelofsen. 2023. [ELAN template for annotation of non-manual markers](#).
- Posit team. 2024. [RStudio: Integrated development environment for R](#). manual, Posit Software, PBC, Boston, MA.
- R Core Team. 2021. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- R Core Team. 2022. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Josefina Safar and Connie De Vos. 2022. [Pragmatic competence without a language model: Other-Initiated Repair in Balinese homesign](#). *Journal of Pragmatics*, 202:105–125.
- William M. Grosse Shyu, Eric and William S. Cleveland. 1992. [Local Regression Models](#). In *Statistical Models in S*. Routledge. Num Pages: 68.
- Felix Sze. 2022. [From gestures to grammatical non-manuals in sign language: A case study of polar questions and negation in Hong Kong Sign Language](#). *Lingua*, 267:103188.
- Suramya Tomar. 2006. [Converting video formats with FFmpeg](#). *Linux Journal*, 2006(146):10. Publisher: Belltown Media.
- James P. Trujillo, Julija Vaitonyte, Irina Simanova, and Asli Özyürek. 2019. [Toward the markerless and automatic analysis of kinematic features: A toolkit for gesture and movement research](#). *Behavior Research Methods*, 51(2):769–777.
- Geoffrey S. Watson. 1964. [Smooth Regression Analysis](#). *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4):359–372. Publisher: Springer.
- Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. [Welcome to the tidyverse](#). *Journal of Open Source Software*, 4(43):1686.
- Hadley Wickham and Winston Chang. 2016. [ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics](#).
- Ulrike Zeshan. 2004. [Interrogative Constructions in Signed Languages: Crosslinguistic Perspectives](#). *Language*, 80(1):7–39.
- Robert Östling, Carl Börstell, and Servane Courtaux. 2018. [Visual iconicity across sign languages: Large-scale automated video analysis of iconic articulators and locations](#). *Frontiers in Psychology*, 9(725).

# An Extension of the NGT Dataset in Global Signbank

Ulrika Klomp<sup>1</sup>, Lisa Gierman, Pieter Manders, Ellen Nauta, Gomèr Otterspeer<sup>1</sup>,  
Ray Pelupessy, Galya Stern, Dalene Venter, Casper Wubbolts,  
Marloes Oomen<sup>1</sup>, Floris Roelofsen<sup>1</sup>

SignLab, University of Amsterdam;

Kloveniersburgwal 48, 1012 CS Amsterdam, the Netherlands;

{u.klomp, l.gierman, p.w.j.manders, e.nauta, g.otterspeer, t.pelupessy, g.stern, d.venter2, c.wubbolts,  
m.oomen2, f.roelofsen}@uva.nl



## Abstract

To support language documentation, linguistic research, and acquisition of Sign Language of the Netherlands (NGT), we are expanding the NGT dataset in the lexical database Global Signbank. Our most prioritized goal is to add ca. 11,000 glosses (entries). We further aim at adding ca. 3,000 example sentences and to provide linguistic information with as many glosses as possible. As for linguistic information, Signbank allows for extensive phonological descriptions of signs, and the addition of multiple senses per sign, which we would like to connect to synsets in the Multilingual Sign Language Wordnet. Additionally, we are recording extra video data: we make multiple videos of the same sign, taken from different angles, and videos with non-manual expressions. Furthermore, we are collecting motion capture data, for improved (automatic) sign language recognition and production in the future. In this paper, we describe how we proceed, the decisions that have been made so far, and future uses of the dataset.

**Keywords:** data collection, Signbank, sign language, NGT, documentation, motion capture

## 1. Introduction

The online lexical database Global Signbank (Crasborn et al., 2018) includes datasets from various sign languages, Sign Language of the Netherlands (*Nederlandse Gebarentaal*, NGT) being one of them. The NGT dataset was composed from 2007 to 2023 (Crasborn et al., 2020) and originated from the need to store and access glosses during corpus annotation work. At the end of 2023, the NGT dataset consisted of ca. 4,100 glosses, where each gloss has its own entry (see Section 2 for more information about entries). The main source of this dataset were annotations within the Corpus NGT (Crasborn,

Zwitserlood and Ros, 2008). In 2023, responsibility for the NGT dataset and for changes in Global Signbank were transferred from the Radboud University Nijmegen to the University of Amsterdam. In 2024, a team of mostly Deaf NGT signers (henceforth: the NGT expert team) was composed to work on the Signbank project at the University of Amsterdam. This project runs till December 2024, and aims at extending the NGT dataset in multiple ways (as outlined in the following sections): 1. adding approximately 11,000 glosses; 2. adding example sentences; 3. adding and systematizing senses;

<sup>1</sup> Inspired by Nyst et al. (2022), we provide drawings of the name signs of our project members (following the author order). Illustrations by Casper Wubbolts.

4. adding more video data; 5. adding linguistic information; 6. collecting (and potentially adding) motion capture data. By expanding the NGT dataset in these ways, we envision to support the documentation of NGT, linguistic research into sign languages, and support learners of NGT. In this paper, we report on the current progress in this project, motivate our decisions so far and discuss potential ways of moving forward.

## 2. Adding Entries

Let us first go into the most significant extension of the NGT dataset; the addition of new glosses. Every gloss receives its own entry. With an entry, we mean a gloss with the video, its meanings and all additional information, visible in one webpage – see Figure 1 below.

When signs are encountered in corpus data and do not have an entry in the NGT dataset in Signbank yet, it is relatively easy to gloss them and include them in the database. But the Corpus NGT is no longer being actively annotated. Furthermore, since we aim to add thousands of lexical items in a short period of time, annotating corpus data is not efficient for gaining so many new entries, as annotation work is highly time-consuming in itself. The question then arose: how do we expand the dataset? We decided to let the

Deaf NGT signers in our team be the data source, and document their knowledge of NGT. As inspiration for concepts to add, we are currently using: 1. themed word lists (e.g. on food or crafting); 2. the gloss list from the Flemish SignBank (Vlaams GebarentaalCentrum, 2024); 3. a list of words from the Corpus Spoken Dutch (CGN Version 2.0.3, 2014<sup>2</sup>). We collected about 6,000 potentially useable concepts until now. We are still thinking of efficient ways to collect 5,000 more concepts to reach our goal.

An important decision that was made to get from concept to entry, is that we only collect signs that are used in the Deaf community, instead of developing or making up signs ourselves. The main reason for this, is that we want to document the language as it is. Thus, if the team has not found an existing NGT sign for a certain concept, the concept is then removed from our list of potentially new entries, and thus not included in the database at this point. Originally, we made the decision that multiple people from the NGT expert team should know a certain sign before it could be included, but this was strikingly unworkable – multiple team members experienced that they were often the only team member who used a specific variant of a sign, due to the signers' different linguistic backgrounds (e.g. different schools and ages).

The screenshot shows the entry for the gloss 'DEAF-B' in the NGT dataset. The interface includes a video player with two frames of a signer, and a detailed metadata panel on the right. The metadata panel is organized into sections: 'Senses', 'Annotation instructions', 'Word class', 'Morphology', 'Phonology', 'Minimal Pairs', 'Semantics', 'Relations to Other Signs', 'Relations to Foreign Signs', 'Publication Status', 'Notes', and 'Other Media'.

Lemma ID Gloss	Dutch: DOOF-B
Annotation ID Gloss (Dutch)	DOOF-B
Annotation ID Gloss (English)	DEAF-B
Senses	1. Dutch: doof English: deaf
Annotation instructions	-
Word class	-
<b>Morphology</b>	
<b>Phonology</b>	
Handedness	1
Strong Hand	N
Location	Ear
Contact Type	Final
Relative Orientation: Movement	Finger tips
<b>Minimal Pairs</b>	
<b>Semantics</b>	
Relations to Other Signs	
Relations to Foreign Signs	
Publication Status	
Notes	
Other Media	

Figure 1: The entry of the gloss 'DEAF-B' in the NGT dataset on Global Signbank, with the phonological panel opened to show specifications (Crasborn et al., 2020).

<sup>2</sup> <http://hdl.handle.net/10032/tm-a2-k6>



To give the team more space, and at the same time guard that not (too many) idiosyncratic forms would be included, we therefore decided that a sign should be used by at least one team member and that this team member should know of at least one other deaf signing person that uses this sign. This did not only speed up the process of deciding upon signs that could be added to the NGT dataset, but was also more in line with the composition of the first dataset, where glosses came from signs that were simply encountered in the Corpus NGT data – sometimes only signed by one signer – and then added. This approach might still change in the future, or might be complemented with other collection projects. An example of an approach that we could take inspiration from for future work is to use an app like SignHunter, as described by Hanke et al. (2020).

When we create a new entry, we add the following information: Annotation gloss IDs, Lemma IDs, senses (possible meanings), a quick webcam recording of the sign and basic phonological information. For the NGT dataset, the decision has (previously) been made to use meaningful Annotation ID glosses (vs. a meaningless code or number, for example), where the ID gloss represents a common meaning. Soon after creating this entry, we expand the senses and (other) linguistic information, and replace the quick webcam video with a high-quality video, made in a professional recording studio. The video of the sign shows what we call a “phonological form”, which represents the manual sign without mouthings, body movements or facial expressions. This is done so that the same manual form always receives the same Annotation ID gloss, even if the form has multiple, very different meanings (see also Section 4 below). The form is therefore articulated in the most neutral way. Since one phonological form can easily represent multiple concepts, we make sure the phonological forms of the proposed signs are not already in the database – perhaps under a different gloss than expected (see Section 6 on how to search for a phonological form). To clarify, it is important here that the intended *phonological form* is not represented in the database yet – the *meaning*, however, may be represented by another form. For example, a commonly used sign for the Dutch island ‘Texel’ refers to (the wool of) sheep. When the NGT expert team considers to add this sign to the database, we would first search for ‘Texel’ as a sense in the NGT dataset. We would then see that this sense is not in the database yet, meaning that the concept is not covered in the dataset. However, when we search for the phonological specifications, we see that the sign is already there, under the gloss SHEEP.

We then add the sense ‘Texel’ to this phonological form, and do not make a new entry.

It is also possible, and even desirable for the purposes of our project, to add multiple signs for one concept, as variants. These different variants are likely to receive similar Annotation ID glosses, but are distinguished by different suffixes. For instance, the signs for ‘dog’ in NGT are currently represented by three different manual forms, with the Annotation ID glosses DOG-A, DOG-B and DOG-C (see Figure 2a, 2b and 2c, respectively).



Figure 2a, 2b, 2c: The signs DOG-A (left), DOG-B (middle), DOG-C (right) in the NGT dataset in Global Signbank (Crasborn et al., 2020)

At the moment of writing<sup>3</sup>, we added 1,600 glosses. One can imagine this process of going from a concept to a full-fledged entry is quite time consuming. One team member therefore developed the signCollect platform in which the work is automatized as much as possible (see Otterspeer, Klomp and Roelofsen (accepted) for a more elaborate explanation of this system). Through this platform, the team can propose glosses, keep track of signs that need consultation, check who will record the signs, make professional recordings and save everything together. We therefore expect to be able to speed up the process and to need less time for the next additions.

### 3. Adding Example Sentences

To provide use-in-context information, we will create example sentences for at least 3,000 glosses. Each example sentence will be accompanied by a gloss-by-gloss representation and a Dutch translation. These sentences will be linked to all the glosses it contains. Some of the 3,000 sentences will be developed in collaboration with a different project, where natural sentences for learners of NGT will be created with help of NGT teachers and parents of deaf children. To join forces, our team supports the recording and annotation process of these sentences, after which we may publish applicable sentences on the Signbank website. Other sentences will be taken from the Corpus NGT (Crasborn et al., 2008; Crasborn et al., 2015). Where possible, the original corpus fragment will

<sup>3</sup> April 4, 2024

be included on the Signbank website; otherwise, the sentence will be refilled.

#### 4. Adding and Systematizing Senses

When Global Signbank was developed, it included a functionality to add translation equivalents or keywords to a gloss (Cassidy et al., 2018). In 2023, keywords have been replaced by senses. A sense is a conceptual meaning and signs may easily have multiple senses – either because multiple distinct meanings are involved (as in homonyms), or because several related concepts apply (as in polysemes). The senses can then be grouped so that senses with a similar meaning are mentioned together. The change of providing (groups of) senses instead of keywords, has, however, not systematically been executed for the NGT dataset. Additionally, many English translations of the senses are still lacking. We therefore have several goals for the upcoming year: 1. add Dutch and English senses to the new and existing glosses; 2. systematically group the senses per concept; 3. connect the senses to synsets in the Multilingual Sign Language Wordnet (Bigeard et al., 2022).

So far, for the entries that also had English translations of the senses already, we checked the translation and regrouped the senses when necessary. For example, the Dutch/English groups of senses that are now available for the gloss PT:down (point down, see Figure 3) are: 1. in/in; 2. nu/now; 3. hier/here; 4. zuid/south; 5. daar/there; 6. dit/this. For every new gloss that we add, we immediately add the most salient sense in Dutch and English. We are developing guidelines to add Dutch and English senses and to group them systematically. Apart from the senses that we added to the new glosses, we added approximately 200 senses to already existing glosses.



Figure 3: The sign 'PT:down' in the NGT dataset on Global Signbank (Crasborn et al., 2020).

#### 5. Adding More Video Data

So far, every entry has one video of the sign connected to the Annotation ID gloss, and one picture. The picture is usually the automatically derived middle frame of the video. In the video, the focus is on the plain articulation of the manual form without non-manual expressions (e.g. facial expressions, mouth actions) (see also Section 2 above). Since these plain signs are considered very unnatural, we aim to add one to three videos per entry where facial expressions and/or mouth actions are included in a natural way. These videos are not meant as replacements for the plain signs and they will not receive their own Annotation ID gloss. Instead, they should be seen as additional material that exemplifies possible natural articulation forms of this basic phonological form.

Both the neutral phonological video and the videos with possible articulation forms are recorded with three different cameras, to provide visual information from three different angles. The different angles will help human recognition of the sign, particularly if a handshape is difficult to perceive from the front angle, but can also be used to train automatic recognition by artificial intelligence. In our current set-up, one camera is situated to have the standard front perspective (similar to the perspective in Figure 3). The other two cameras are in a ca. 25-degree angle from the signer on the left and right of the middle camera, as we discovered these are optimal camera positions to capture multiple perspectives.

#### 6. Adding Linguistic Information

Global Signbank allows for extensive description of linguistic information on different levels for every glossed sign – although the different datasets in Signbank vary in the extent to which they make use of these possibilities. For the NGT dataset, it has been a specific goal to collect phonological information (Cassidy et al., 2018) and therefore the possibilities to describe phonological characteristics of signs are quite elaborate. For each entry, one can fill out several fields on handshape(s), location, movement, orientation and, if necessary, other additional information about the sign. See, as an example, Figure 1 for the phonological description of the sign DEAF-B.

The description of phonology is mostly done through the selection of features in drop-down menus, to make the process easier, more standardized and less prone to typos. An advantage of this standardization is that it makes it easier to look up whether a phonological form is already in the dataset. When looking for a phonological form, one can fill out the relevant phonological information and find any relevant

sign without having to know the possible senses of the sign. Furthermore, Global Signbank allows for automatic searches for minimal pairs, for which the phonological information is used.

Note that, which phonological information is considered relevant, is also depending on the theoretical framework one is working with. The current structure of the phonology fields in Global Signbank reflects the line of work performed and followed at the Radboud University Nijmegen – and now by our team –, i.e., based on the work of e.g. Crasborn (2001) and van der Kooij (2002). At the moment of writing, we added phonological information for the majority of the 1,600 newly added signs.

Another section in the detailed view of an entry is related to morphological information, where one can describe if a sign is a compound, and if yes, what the individual compounded parts are. Within our project, we will add phonological information on the newly added glosses, and potentially investigate possibilities to describe compounds more elaborately. We will also look into the descriptions made by other datasets in Global Signbank, to enhance comparability among the datasets.

## 7. Collecting Motion Capture Data

To enhance and support developments in automatic sign language recognition and production, we are collecting motion capture data. By collecting data from the same signers and on the same signs that we collect for the NGT dataset, we create a big dataset with datapoints from different types (2D video data, 3D motion capture data) that all relate to the same concepts. In our current set-up, we use 12 infrared cameras, most of which are located on the ceiling to record from above, and a few on the ground to record from below (see Figure 4). We use the motion capture suit of Vicon, where we reconstruct a scene in 3D through the markers on this suit. Facial movements are captured with Live Link of Unreal, supported by ARKit of Apple. Additionally, we use StretchSense gloves to measure hand and finger movements (position and configuration of the hand and fingers). In Figure 4, one of our team members is preparing to produce the sign presented on the left screen, while wearing the motion capture suit and the StretchSense gloves. The screen on the right in Figure 4 shows an avatar, reconstructed from the signer in real time.

Processing of the data is done with Unreal Editor 5.3 to combine the data stream in a so-called FBX file. We use the signCollect system (Otterspeer, Klomp and Roelofsen, accepted) for directing the systems, collecting, saving and labelling the data. We are still practicing and experimenting with this set-up, but the results so far are promising: we have been able to record 1,000 glosses in this set-up by now. If it seems useful, the motion capture

data will also be added to the NGT dataset in Signbank.



Figure 4: The set-up for recording motion capture data for lexical signs.

## 8. Future Directions

Global Signbank already has the possibility of performing automated searches and basic analyses. It is, for example, possible to automatically look for minimal pairs, or provide a distribution of the most frequently occurring handshapes. The more data in the NGT set, and the more precise their description, the more reliable these outcomes will be. Additionally, if one has access to multiple datasets, one can easily make cross-linguistic comparisons with these tools. Thus, expanding the NGT dataset supports linguistic research.

The original NGT dataset has frequency data for occurrence of the signs in the Corpus NGT available. In future research, we would like to collect frequency data on newly added signs as well (see e.g. Johnston (2012) on why lexical frequency data is relevant) – either by taking frequencies from the corpus, or by eliciting frequency measures from a large group of Deaf NGT signers.

With the extension of the NGT dataset, it will also be a richer platform for learners of NGT. The addition of signs, senses, examples sentences and videos from different angles support in acquiring a rich vocabulary and in understanding the different meanings a sign may have. The database could at some point also function as a dictionary. This is important, because not many sign language dictionaries exist for NGT. Furthermore, Signbank is freely accessible, and allows for searching from Dutch or English words (senses) to signs, but also the other way around, by searching with the phonological specifications.

Lastly, the video data and motion capture data will be used for automatic recognition and production of sign languages. By providing language models with our extensive dataset, we support the development of automatic translations from written language to sign language and vice versa.



## 9. Acknowledgments

This research is funded by Platform Digital Infrastructure Social Science and Humanities (PDI-SSH).

## 10. Bibliographical References

Bigear, S., Schulder, M., Kopf, M., Hanke, T., Vasilaki, K., Vacalopoulou, A., Goulas, T., Dimou, A.-L., Fotinea, S.-E., and Efthimiou, E. (2022). Introducing Sign Languages to a Multilingual Wordnet: Bootstrapping Corpora and Lexical Resources of Greek Sign Language and German Sign Language. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France. European Language Resources Association (ELRA).

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E., and Johnston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2359–2364, Miyazaki, Japan. European Language Resources Association (ELRA).

Crasborn, O. (2001). *Phonetic implementation of phonological categories in Sign Language of the Netherlands*. PhD dissertation, Leiden University.

Crasborn O., Bank, R., Stoop, W., Komen, E., Hulsbosch, M., and Even, S., (2018). *Global Signbank source code*. Radboud University, Nijmegen.

Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., Schüller, A., Ormel, E., Nauta, E., van Zuilen, M., van Winsum, F., and Ros, J. (2016). Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT. In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 41–46, Portorož, Slovenia. European Language Resources Association (ELRA).

Hanke, T., Jahn, E., Wähl, S., Böse, O., and König, L. (2020). SignHunter – A sign elicitation tool suitable for deaf events. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages*, pages 83–88, Marseille, France. European Language Resources Association (ELRA).

Johnston, T. (2012). Lexical frequency in Sign Languages. *Journal of deaf studies and deaf education*, 17(2):163–193.

van der Kooij, E. (2002). *Phonological categories in Sign Language of the Netherlands: The role of phonetic implementation and iconicity*. PhD dissertation, Leiden University.

Nyst, V., Morgado, M., Mac Hadjah, T., Nyarko, M., Martins, M., van der Mark, L., Burichani, E., Angoua, T., Magassouba, M., Sylla, D., Admasu, K., and Schüller, A. (2022). Object and handling handshapes in 11 sign languages: towards a typology of the iconic use of the hands. *Linguistic Typology*, 26(3):573–604.

Otterspeer, G., Klomp, U., and Roelofsen, F. (Accepted). SignCollect: A ‘touchless’ pipeline for constructing large-scale sign language repositories. To appear in: *Proceedings of the LREC2024 11th Workshop on the representation and processing of sign languages: Evaluation of sign language resources*.

## 11. Language Resource References

Corpus Gesproken Nederlands - CGN (Version 2.0.3). 2014. Data set. Available at the Dutch Language Institute: <http://hdl.handle.net/10032/tm-a2-k6>

Crasborn, O., Bank, R., Zwitserlood, I., van der Kooij, E., Ormel, E., Ros, J., Schüller, A., de Meijer, A., van Zuilen, M., Nauta, E., van Winsum, F., and Vonk, M. 2020. Nederlandse Gebarentaal (NGT) dataset in Global Signbank. In O. Crasborn et al., (Eds.) *Global Signbank*. Radboud University, Nijmegen. ISLRN: 976-021-358-388-6, DOI: 10.13140/RG.2.1.2839.1446.

Crasborn, O., Zwitserlood, I., and Ros, J. 2008. *Het Corpus NGT. Een digitaal open access corpus van filmpjes en annotaties van de Nederlandse Gebarentaal [The Corpus NGT. A digital open access corpus of movies and annotations of Sign Language of the Netherlands]*. Nijmegen: Centre for Language Studies, Radboud University.

Vlaams GebarentaalCentrum. 2024. VGT Signbank [dataset]. <https://vlaamsegebarentaal.be/signbank>.



# Corpus à la carte – Improving Access to the Public DGS Corpus

Reiner Konrad , Thomas Hanke , Amy Isard , Marc Schulder ,  
Lutz König , Julian Bleicken , Oliver Böse 

Institute of German Sign Language and Communication of the Deaf

University of Hamburg, Germany

{reiner.konrad, thomas.hanke, amy.isard, marc.schulder,  
lutz.koenig, julian.bleicken, oliver.boese}@uni-hamburg.de

## Abstract

This article presents the fourth release of the Public DGS Corpus, a large corpus of German Sign Language (DGS). Since its first release in 2018, the Public DGS Corpus has provided its content through multiple portals to meet the needs of different user groups. Having started with a community portal and a research portal for general data access, the ANNIS portal for dynamic web-based exploration of the corpus was added in 2022. With this latest release, a fourth portal is added to allow sign language linguists to access the public corpus directly through the annotation software *iLex*. Furthermore, search capabilities and interconnectedness between the portals are strongly improved, allowing users to move between portals to combine their strengths. Additional improvements to the corpus include additional recordings, new pose information models, improved HamNoSys, enhanced type information and web interface revisions.

**Keywords:** German Sign Language (DGS), Linguistic Resource, Corpus, Resource Extension

## 1. Introduction

This paper presents the fourth release of the Public DGS Corpus, introducing new features and content as well as a new portal, *MY DGS – iLex*. It is a follow-up of Jahn et al. (2018) and Hanke et al. (2020), which described the previous releases of the corpus. A special focus is given to demonstrating how interconnectedness between the different corpus portals helps improve access to the data of the Public DGS Corpus. The paper also introduces the upcoming second data collection phase of the DGS Corpus, in which new primary materials will be recorded.

The paper is structured as follows: Section 2 gives a brief description of the DGS-Korpus project, the upcoming second data collection, and the history of releasing the Public DGS Corpus through a set of portals optimised for different use cases. The new features and content of release 4, including the new portal *MY DGS – iLex*, are presented in Section 3. Section 4 describes the connections between the portals, and Section 5 works through two use case examples which illustrate how connections between the portals can be used to combine the different strengths of each portal to match the user's needs, making the Public DGS Corpus a "corpus à la carte".

## 2. The DGS Corpus

The DGS-Korpus project (2009–2027) is a long-term research project to create a reference corpus of German Sign Language (DGS; *Deutsche Gebärdensprache*) (Prillwitz et al., 2008). Building on the DGS Corpus, the project has also produced a corpus-based dictionary, *DW-DGS* (Langer et al., 2024) and the Public DGS Corpus dataset which is the focus of this article. The purpose of the corpus is to be both a resource for linguistic research and a record of deaf heritage in Germany.

*densprache*) (Prillwitz et al., 2008). Building on the DGS Corpus, the project has also produced a corpus-based dictionary, *DW-DGS* (Langer et al., 2024) and the Public DGS Corpus dataset which is the focus of this article. The purpose of the corpus is to be both a resource for linguistic research and a record of deaf heritage in Germany.

### 2.1. Data Collection Phases

The majority of data for the DGS Corpus was gathered during its first data collection phase (2010–2012), during which dyadic conversations between 330 participants from thirteen regions of Germany and four different age groups were recorded, resulting in 1150 hours of recordings, containing 560 hours of semi-spontaneous DGS signing.

A second data collection is scheduled for 2024–2025 to add a fifth cohort of 46 participants, most of whom will be aged 18–32, i.e. people who have come of age since the first collection phase. Performing this second collection with a younger cohort allows the corpus to cover further relevant developments affecting the German deaf community, such as changes in educational policy (progressing from bilingual pilot projects to integration and then to inclusivity), changes in information technology (use of smart phones, social media, video telephony), medical advances in the use of cochlear implants, demographic changes, international mobility and language contact.

The second data collection will follow the design of the first collection, with a number of updates to account for changes in technology and participant

background. New camera equipment will be used, increasing video resolution from 2K to 6K. The collection formats will stay the same, but some tasks and stimuli are updated. For example, the list of historical events has been changed to match the time frame actively experienced by participants, a new discussion topic regarding social media has been introduced, and various images have been switched out to reference more recent politicians, celebrities and events as well as contemporary technology hardware (e.g. replacing images of CRT monitors with flatscreens).

## 2.2. The Public DGS Corpus: Thinking with Portals

A part of the DGS Corpus was selected for inclusion in a fully annotated publicly available dataset, the Public DGS Corpus.<sup>1</sup> The Public DGS Corpus was initially released in 2018 (Jahn et al., 2018) and extended in content and features through subsequent releases in 2019 and 2020 (Hanke et al., 2020). To accommodate the needs of different user groups (see Jahn et al., 2018), multiple web portals were created:

**MY DGS** is a community portal for the deaf community, presenting the corpus as a heritage resource, with a focus on easily finding interesting conversations about various aspects of deaf culture and life experience. To accommodate language learners and others interested in deaf culture, optional German subtitles are provided.

**MY DGS – annotated** targets linguistic researchers, providing all recordings with full annotations and translations in German and English, available for download and display through an online viewer. It also provides machine-readable metadata files and pose information for computational processing. In addition to the recordings of the community portal, some further recordings are provided that cover tasks with purely linguistic, rather than cultural, value, such as retellings of stories.

**MY DGS – ANNIS** has been available as a third portal since 2022, as described in Isard and Konrad (2022). ANNIS is a web-based corpus search tool which allows users to search and visualise corpus data (Krause and Zeldes, 2016). The portal integrates the Public DGS Corpus into an ANNIS instance to allow dynamic exploration of corpus contents without the need for installing annotation software.

**MY DGS – iLex** is a new portal introduced with release 4. It provides the content of the Public DGS Corpus as a relational database that can be accessed via iLex (Hanke, 2002), a tool for lexicographic and corpus linguistic research.

<sup>1</sup>For a discussion of why only a part of the DGS Corpus is made public, see Schulder and Hanke (2022).

## 3. Changes in Release 4

In this section we describe the new features, functions and kinds of information made available by release 4 of the Public DGS Corpus. Section 3.1 describes the new corpus transcripts introduced by the release, which are added to all portals. The remaining descriptions are grouped by portal, starting with updates to *MY DGS* (Section 3.2) and *MY DGS – annotated* (Section 3.3), moving to how release 4 is integrated into *MY DGS – ANNIS* (Section 3.4) and a description of the new portal *MY DGS – iLex* (Section 3.5).

### 3.1. New transcripts

Release 4 introduces one hour of additional material to the Public DGS Corpus, bringing its full size to 52.4 hours. *MY DGS – ANNIS*, which excludes 2.4 hours of videos which have no annotations (see Isard and Konrad, 2022), grows from 49 to 50 hours.

The new material provides four retellings (7 minutes) and 18 process descriptions (53 minutes). The retellings are fully translated and lemmatised. They include two pear story retellings (Chafe, 1980) and two retellings of the broadcast “The Domestic Aid” (Sehen statt Hören, 2006). The new process descriptions are provided with translations, but without lemmatisation. They cover processes such as preparing a meal, baking a cake, or mending a puncture.

*MY DGS* does not cover certain tasks that are considered mainly of interest to linguistic research. For this reason, the new retelling recordings are omitted. However, for release 4 it was decided that process descriptions should be part of *MY DGS*, so the 18 new process descriptions are added, as well as the 13 process descriptions (42 minutes) that had been added to *MY DGS – annotated*. This brings the total size of *MY DGS* to 51.3 hours.

### 3.2. Changes to MY DGS

Release 4 of *MY DGS* offers enhanced search functionality to make content discovery even easier. In addition to filter options for region, age group, conversation format, and topics of conversation, a new full-text search on the German translations is introduced.

This search narrows down the selection of videos to those containing the searched text in their subtitles. When a specific video is selected, a list of all matching subtitle lines is shown, including the timecode at which it appears in the video and a button to start the video at that timecode. An example of this is shown in Figure 1. Text search also supports the wildcard symbol “\*”, so as a side effect users can generate a full text transcript of the German

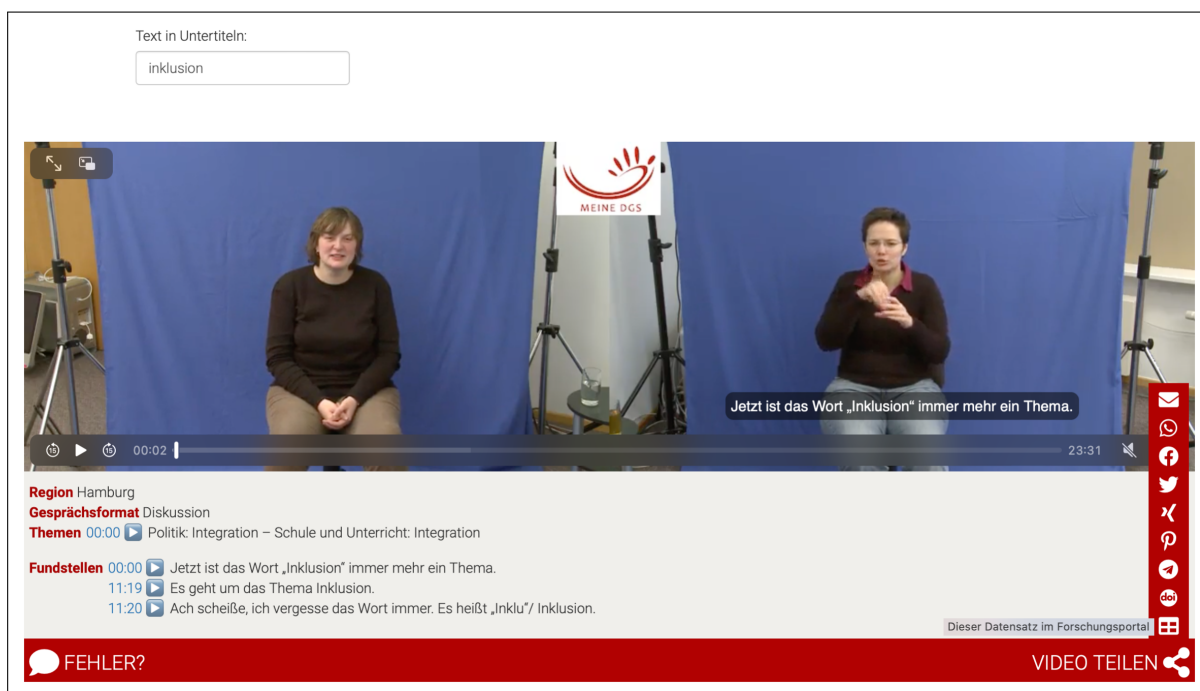


Figure 1: *MY DGS* video with subtitle search field above the video and metadata (region, format, topic) and subtitle search results below. In the lower right corner are buttons for jumping to *MY DGS – annotated* and sharing via social media or email.

translations by entering only the wildcard by itself in the search field.

A new search field was also added to the topics filter (“Alle Themen”) which now lists, in addition to the 33 main topics, more than 570 keywords (previously only included in *MY DGS – annotated*). Browsing keywords and using the search function helps to quickly find videos of interest.

### 3.3. Changes to *MY DGS – annotated*

*MY DGS – annotated* receives improvements to its web interface and the amount of information and data it provides.

#### 3.3.1. Type Entries

The pages for individual type entries have received several improvements, particularly regarding information provided for subtypes, i.e. specific meanings of a sign (see “double glossing” in Konrad et al. (2022) for details). An example of the subtype section of a type entry can be seen in Figure 2.

Where a mouthing typically accompanies a sign meaning, this is now specified in the subtype entry. Where available, this is accompanied by a video of the sign’s citation form with the appropriate mouthing, which functions as a supplement to the citation form video without mouthing that is provided for the main type.

Subtypes now also provide translational equivalents for the given meaning of a sign. Depend-

ing on the chosen interface language, these are translations into German or English. Translational equivalents are sourced from the lexical inventory of the DGS Corpus iLex database and are specified based on the complete reference corpus as well as other information sources, so they may include translations that are not based on tokens of the Public DGS Corpus.

The HamNoSys notations provided for sign types have undergone a major quality assurance revision, improving their quality and consistency.

At the bottom of the type entry page, a table of downloadable data for the type and its subtypes is provided. Similar to the download table for transcripts, it provides annotation files in iLex XML, ELAN and SRT formats, video files of the available camera perspectives, and pose information.

#### 3.3.2. Types List

The types list receives a new search feature that lets users filter the list of type glosses by entering (partial) gloss strings or translational equivalents. The search also allows filtering by phonetic attributes via HamNoSys notation.<sup>2</sup> Gloss and HamNoSys strings can be combined to narrow down the search further, as can be seen in Figure 3.

<sup>2</sup>To produce HamNoSys symbols, one can use the platform independent HamNoSys editor at <https://www.sign-lang.uni-hamburg.de/hamnosys/input/>

=

## INKLUSIV1 (26 Tokens)

Mund: inklusiv

frontal

schräg von vorne

seitlich

von oben

Übersetzungsäquivalente: Inklusion; inklusiv; inklusive

▶ 1178939 hh07 | 31-45f Jetzt ist das Wort „Inklusion“ immer mehr ein Thema.

r	MEHR3	AUCH3A*	WORT3	INKLUSIV1
l				
m	mehr mehr	auch		inklusion

Figure 2: Excerpt of a type entry page in *MY DGS – annotated*, showing the subtype INKLUSIV1. Subtypes now specify their typical mouthing (here: *inklusive*) and provide a mouthing-specific citation form video (when available) and translational equivalents in German or English, depending on the interface language.

In addition to the ordering by type gloss, a new ordering by translational equivalents is provided. As discussed in Section 3.3.1, these are German or English translations for subtypes of a sign. The type list gathers all translations, sorted alphabetically, and for each translation lists all subtypes that specify it as a semantic equivalent.

Subtypes with multiple translational equivalents are listed repeatedly, as can be seen in Figure 4. For instance, INKLUSIV1 is listed for “Inklusion” (inclusivity), “inklusive” (inclusive) and “inklusive” (including), while INKLUSIV2 and INKLUSIV3 have only the latter two as translational equivalents. This types list ordering can also be filtered, in this case based on (partial) strings of translational equivalents.

### 3.3.3. Transcripts

The transcript overview has been extended to provide the numeric identifier of each transcript that are used in corpus filenames and DOIs, making it easier for users to find the correct transcript for a downloaded file.

The overview also received a visual indicator for whether lemmatisations, translations or only video material is available for a given transcript.

The transcript viewer has been updated to support the connections to other portals described in Section 4.

### 3.3.4. Data Collection Formats

The pages describing individual data collection formats now provide a link to the corresponding entry in the *Sign Language Dataset Compendium* (Kopf et al., 2022), a resource compiling information on corpora and lexical resources for sign languages, including a list of commonly used collection tasks and which corpora include them.

### 3.3.5. Pose information


*MY DGS – annotated* provides pose information for computational analysis of corpus data. Pose information represents participants as a set of automatically determined keypoints, image coordinates of specific points on the body. Until now, all pose information provided by *MY DGS – annotated* was generated using the OpenPose pose recognition model (Cao et al., 2021; Simon et al., 2017) (see Hanke (2019) and Schulder and Hanke (2020) for details). With release 4, additional outputs generated using MediaPipe (Lugaresi et al., 2019) and Apple Vision Framework<sup>3</sup> pose models are made available. As a result, several alternative pose representations are now available for every corpus recording, including representations with 3D estimates for keypoint locations.

<sup>3</sup><https://developer.apple.com/documentation/vision/>



nach Glossen | nach Translationsäquivalenten


**Filter:**

inklu 

INKLUSIV1^ (46 Tokens)

INKLUSIV1 (26 Tokens) → INKLUSIV1^

INKLUSIV2 (5 Tokens) → HINEINSTECKEN2^

Figure 3: *MY DGS – annotated* gloss search that combines text string search with a phonetic restriction to only show glosses whose sign type uses a hand with fingers that are pinched together while stretched out, signified by HamNoSys notation .

### 3.4. Changes to MY DGS – ANNIS

The *MY DGS – ANNIS* portal has been available since 2022, and allows researchers to search the German or English version of the Public DGS Corpus (Isard and Konrad, 2022). Queries can cover multiple transcripts and/or multiple annotation tiers, and can also include corpus metadata. Search results are shown with annotations displayed as a horizontal grid, linked to the video file for a transcript. Export of results in csv format is possible if further processing of the results is required.

Release 4 of the Public DGS Corpus is made available through *MY DGS – ANNIS*, but access to the release 3 data is maintained, as is the case with subsequent Public DGS Corpus releases through other portals. For a demonstration of the *MY DGS – ANNIS* interface, see Figure 6 and Section 5.

With release 4, we add new features to *MY DGS – ANNIS*, including an additional tier of keywords linked to the existing tiers, so it is possible for example to search for vocabulary which occurs particularly often during discussions of a particular topic. Release 4 also introduces links to the other Public DGS Corpus portals as described in Section 4.

*MY DGS – ANNIS* has proved to be a popular resource among the sign language research community, but the richness of the annotation data combined with the need to learn the basics of the query language AQL can be a barrier for some researchers. We have therefore introduced a Query Wizard which allows users to build up a query using an interface which requires basic knowledge of the annotations of the Public DGS Corpus, but no prior knowledge of AQL. A detailed description of the Query Wizard can be found in Isard (2024). The Query Wizard is compatible with each *MY DGS – ANNIS* dataset (each covering a specific corpus release and annotation language) and will be updated for use with future corpus releases. This has the

Inhalt	INHALT1 (27 Tokens) → HINEINSTECKEN2^ INHALT3 (134 Tokens) → HINEINSTECKEN3^
Inklusion	INKLUSIV1 (26 Tokens) → INKLUSIV1^
inklusiv	INKLUSIV1 (26 Tokens) → INKLUSIV1^ INKLUSIV2 (5 Tokens) → HINEINSTECKEN2^ INKLUSIV3 (3 Tokens) → HINEINSTECKEN3^
inklusive	INKLUSIV1 (26 Tokens) → INKLUSIV1^ INKLUSIV2 (5 Tokens) → HINEINSTECKEN2^ INKLUSIV3 (3 Tokens) → HINEINSTECKEN3^
Inkorporation	INKORPORATION1 (4 Tokens) → INTEGRATION1^
Inliner	INLINER1A (5 Tokens) → SCHLITTSCHUH- LAUFEN1B^

Figure 4: Excerpt of *MY DGS – annotated* type list, grouped by translational equivalent. Subtypes with multiple possible translations are repeated for each of them.

benefit of allowing researchers to seamlessly transition between versions, even when structural differences between the datasets necessitate changes to the AQL queries.

### 3.5. Introducing MY DGS – iLex

Users with experience in using the annotation software iLex can now use it to access a read-only version of the Public DGS Corpus through the *MY DGS – iLex* portal. This representation most closely matches the internal annotation environment of the DGS-Korpus project. As such it provides strong support for advanced structures of the corpus, such as the type hierarchy and double token tags (Konrad et al., 2022).

Users familiar with interfacing with PostgreSQL databases can also directly access the iLex database of the *MY DGS – iLex* portal.

## 4. Connecting the Portals

The Public DGS Corpus is a resource for a variety of groups, such as the deaf community, linguistic researchers and sign language educators and students. To serve these different groups, different corpus portals were optimised for the needs of specific groups and their use cases. Yet as the corpus grew and evolved, it became clear that the interests of the individual groups were not necessarily limited to a single portal. Rather, the optimised experience of one portal could lead users to become familiar enough with its contents to wish to explore additional facets better served by another portal.

For instance, *MY DGS – annotated* was initially intended as purely targeting the international research community and therefore its interface was available only in English. However, feedback from

From	To	Link
<i>MY DGS</i>	<i>MY DGS – annotated</i>	to the same transcript in the other portal
<i>MY DGS – annotated</i>	<i>MY DGS</i>	to the same transcript in the other portal
	<i>MY DGS – ANNIS</i>	from each type/subtype token, translation or mouthing annotation to the search result for the corresponding item
	<i>MY DGS – iLex</i>	from each type/subtype token, translation or mouthing annotation to the corresponding iLex entry
<i>MY DGS – ANNIS</i>	<i>MY DGS – annotated</i>	from search results to corresponding position in transcript or entry in the types list
	<i>MY DGS – iLex</i>	from each item in search results to the corresponding iLex type or token.
<i>MY DGS – iLex</i>	<i>MY DGS</i>	to the same transcript
	<i>MY DGS – annotated</i>	from each type/subtype token, translation or mouthing annotation to the corresponding timecode or entry in the types list
	<i>MY DGS – ANNIS</i>	from each type/subtype token, translation or mouthing annotation to the search result for the corresponding item

Table 1: Table of the links between the four Public DGS Corpus portals

the deaf community<sup>4</sup> showed that they were interested in also exploring the research data aspect of the Public DGS Corpus, so in release 2 a German interface was added to provide better accessibility (Hanke et al., 2020).

With release 4, the portals turn from separate resources into a network of interfaces to explore the Public DGS Corpus and be combined for new emerging use cases. This is a continuation of our interconnectivity efforts that started with the inclusion of links to DW-DGS and several specialist dictionaries on type entry pages in release 3.

Depending on the portal, different kinds of links to the other portals exist. These connections are described in Table 1. An illustration of the network of connections can also be seen in Figure 5.

Users can move from within a transcript to the same transcript in other portals. Most commonly this leads to the beginning of the transcript, although depending on the structure of the in- and outgoing portals, jumping to a specific timecode or token might also be possible. Connections for types are also available. These lead either to type entries of the other portal or, in the case of *MY DGS – ANNIS* to a query listing all token occurrences of the type.

For some concrete examples of how the interconnected portals can be used, see the use cases described in Section 5.

<sup>4</sup>Such feedback was received via the project’s dedicated “focus group”, a team of deaf experts providing advisory and outreach capacities that connect the project to communities of different regions (Prillwitz et al., 2008).

## 5. Use Cases

In this section we describe some use case scenarios to illustrate ways to start a search in the Public DGS Corpus and how to profit from the connections between portals.

### 5.1. From topics and keywords to meanings and tokens

Let us assume that you are interested in whether inclusivity was already a topic of debate in the early 2010s. You can search for the keyword “Inklusion” in the topics of *MY DGS* (filter “Alle Themen”) where you will find six videos. Opening a video and using the search field “Text in Untertiteln” (text in subtitles) you will find the German translations containing “Inklusion”. Using the timecode link you can direct the video to the starting point of the translation.

Not all occurrences of inclusivity occur in transcripts which have the topic “Inklusion”, so searching for “inklu” in the subtitles search field will return a larger number of hits, namely twelve videos. The subtitles contain 16 translations with “Inklusion” or the adjective “inklusive” or the verb “inkludieren”, all in the sense of ‘concerning inclusivity’, and 10 translations with the adjective “inklusive” in the English sense of “all inclusive”.

Moving on to *MY DGS – annotated*, you can search the types by translational equivalents<sup>5</sup> in order to find subtypes which are associated with German words beginning with “inklu”. For example,

<sup>5</sup>[https://meine-dgs.de/ling/meanings\\_de.html](https://meine-dgs.de/ling/meanings_de.html)

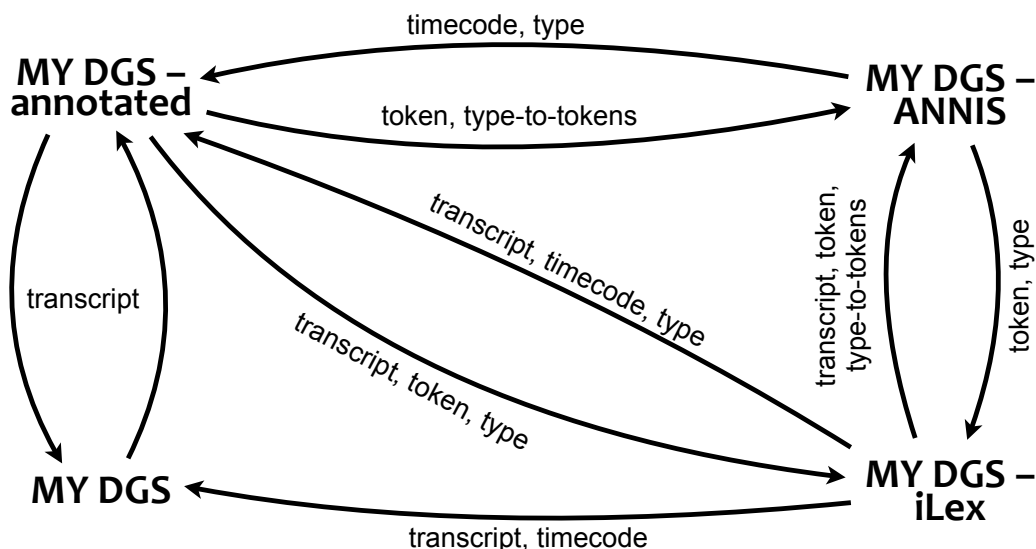


Figure 5: Diagram of the connections between the four Public DGS Corpus portals. *Transcript* denotes links to a given video transcript, *timecode* to a specific time point in the video, *token* to a specific token or mouthing/translation annotation in the transcript, *type* to an entry in the types list, *type-to-token* to a search of all tokens that are instances of the given type.

INKLUSIV<sup>6</sup> has the equivalents “Inklusion; inklusiv; inklusive” (inclusivity; inclusive; included). For each type you can then inspect the translations manually to find the German translations containing “Inklusion” or the adjective “inklusiv” in the sense of ‘concerning inclusivity’, but not in the sense of ‘all inclusive’. From the type entry you can also use the links to *MY DGS - ANNIS* and *MY DGS - iLex* to continue and refine your search.

In *MY DGS - ANNIS*, you can also use AQL query expressions to look for links between different annotation tiers. Instead of looking at the type entry, you could, for example, use [Query 1](#) to search for translations with “inkl” linked to glosses which have a Mouthing which also starts with “inkl”. This returns 19 matches, as shown in [Figure 6](#).

- (1) Gloss ->ident Mundbild=/inkl.\*/ & Deutsch=/.\*[il]nklu.\*/ & #1 ->ident #3

From the matches in *MY DGS - ANNIS* you can jump to either the type entries or the relevant transcript sections of *MY DGS - annotated* and *MY DGS - iLex*, as described in [Section 4](#).

In *MY DGS - iLex* you can use a customized SQL query searching the German translations for “inkl” but excluding “alles inklusiv” and “inklusive alle”. You can also make use of a number of predefined SQL functions of *MY DGS - iLex*, such as the `tag_to_glossstring()` function, which takes a translation tag ID and outputs the sequence of token glosses that are covered by the translation. In our search, this function could be added to the

<sup>6</sup><https://doi.org/10.25592/dgs.corpus-4.0-type-51514#type51515>

SQL query to not only see the relevant translations, but also the textual representation of the underlying signed utterance to make a preliminary confirmation regarding the relevance of the sentence.

## 5.2. Idiomatic phrases

Looking for collocations is one step in the lexical description of words and signs. This may lead to the discovery of multi-word expressions or even idiomatic phrases. The latter seem to be very rare in sign languages ([Johnston and Ferrara, 2012](#); [Wilkinson et al., 2023](#)). Some examples of phrases in DGS can be found in the Digital Dictionary of DGS (DW-DGS)<sup>7</sup>. For example, entry 262<sup>8</sup> identifies the phrase WARM1A<sup>9</sup> GROUP1A<sup>10</sup> which is described as meaning *cordial* or *communal* in the sense of cohesion and interaction in groups of people. The DW-DGS entry provides three examples of the phrase being signed in the DGS Corpus.

Based on the information in DW-DGS, one could assume that “WARM1A^ GROUP1A^” is a fixed phrase with a semantically idiomatic status (cf. [Wilkinson et al., 2023](#)). To confirm this with the Public DGS Corpus data, one can run searches in *MY DGS - ANNIS*. The strength of this corpus search tool is that you can easily define the context of a token,

<sup>7</sup>See [Langer et al. \(2022\)](#) for a detailed description of information types in the DW-DGS.

<sup>8</sup><https://www.sign-lang.uni-hamburg.de/korpusdict/bags/bag262.html>

<sup>9</sup><https://doi.org/10.25592/dgs.corpus-4.0-type-13170>

<sup>10</sup><https://doi.org/10.25592/dgs.corpus-4.0-type-13141>

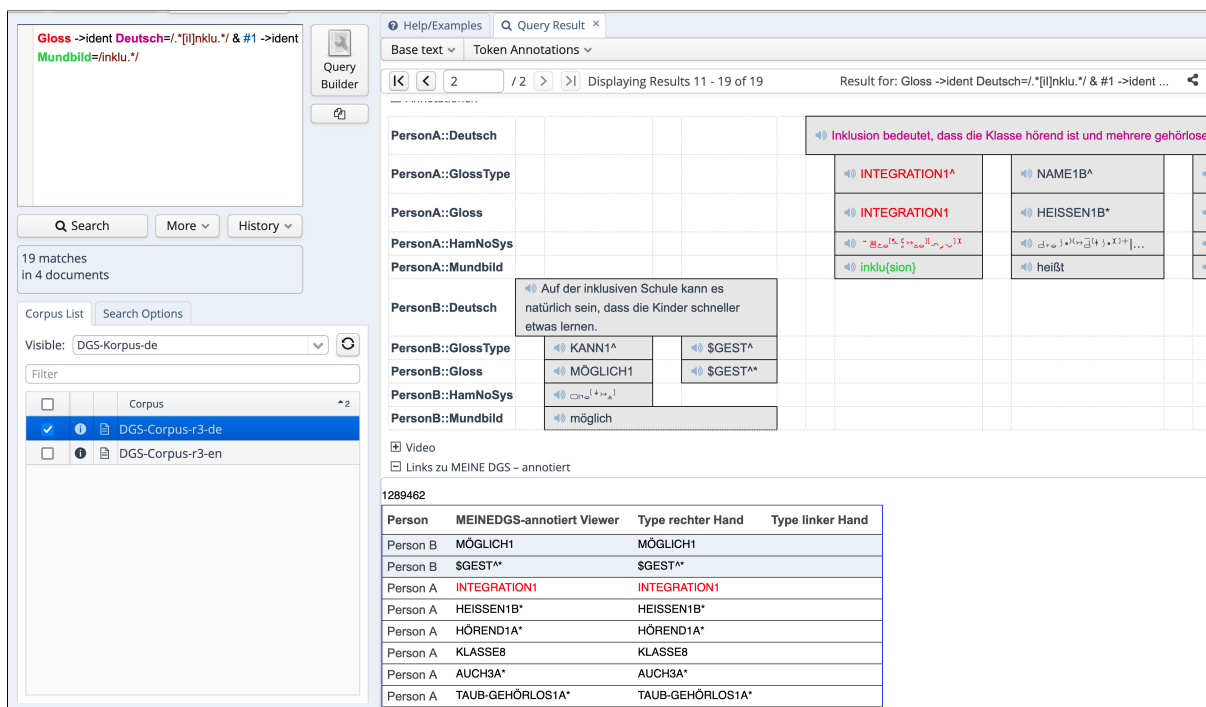


Figure 6: MY DGS – ANNIS search for German translation containing “[li]nkle” and Mundbild including “inklu”.

i.e. the preceding or following neighbours, specifying the distance between the search item and its neighbours. The numbers and letters after the gloss name are not specified in this and the following queries in order to retrieve all combinations of variants of types with the gloss names WARM and GROUP. For details on how you can build these AQL queries with the ANNIS Query Wizard see [Isard \(2024\)](#).

First, you can check the the sequence WARM GROUP using [Query 2](#), for which there are 8 matches in the Public DGS Corpus.

- (2) `GlossType=/WARM.*/.GlossType  
GlossType=/GROUP.*/  
& Gloss ->ident English & #3 ->ident #1`

Second, you can check the reverse sign order, GROUP WARM, using [Query 3](#), for which there is 1 match.

- (3) `GlossType=/GROUP.*/.GlossType  
GlossType=/WARM.*/  
& Gloss ->ident English & #3 ->ident #1`

Third, you can search for co-occurrences of WARM and GROUP in either order and with one to four tokens in between using [Query 4](#), for which there are 10 matches.

- (4) `GlossType=/GROUP.*/^GlossType,2,5  
GlossType=/WARM.*/  
& Gloss ->ident English & #3 ->ident #1`

A closer look at the data has shown us that WARM GROUP is not just a fixed combination, but may also occur in reverse order and interrupted by other lexical elements.

## 6. Conclusion

We presented release 4 of the Public DGS Corpus. It introduces one more hour of recordings from the DGS Corpus, new pose data and a new portal, *MY DGS – iLex*. All portals receive improvements such as added information, new search capabilities and other interface refinements.

A major change in release 4 is the added interconnectivity between portals. Each portal offers a number of ways to jump to other portals, allowing for a more dynamic use of the resources, combining each of their strengths.

## 7. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.



## 8. Bibliographical References

- Zhe Cao, Ginés Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A. Sheikh. 2021. [OpenPose: Realtime multi-person 2D pose estimation using part affinity fields](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186.
- Wallace L. Chafe, editor. 1980. *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Number 3 in *Advances in discourse processes*. Ablex Publishing Corporation, Norwood, New Jersey, USA.
- Thomas Hanke. 2002. [iLex - a tool for sign language lexicography and corpus analysis](#). In *3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).
- Thomas Hanke. 2019. [Processing DGS-Korpus data with OpenPose on the Hamburg High Performance Cluster](#). Project Note AP04-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. [Extending the Public DGS Corpus in size and depth](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).
- Amy Isard. 2024. [Building your query step by step: A Query Wizard for the MY DGS – ANNIS portal of the DGS Corpus](#). In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy. European Language Resources Association (ELRA).
- Amy Isard and Reiner Konrad. 2022. [MY DGS – ANNIS: ANNIS and the Public DGS Corpus](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 73–79, Marseille, France. European Language Resources Association (ELRA).
- Elena Jahn, Reiner Konrad, Gabriele Langer, Sven Wagner, and Thomas Hanke. 2018. [Publishing DGS Corpus data: Different formats for different needs](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 83–90, Miyazaki, Japan. European Language Resources Association (ELRA).
- Trevor Johnston and Lindsay Ferrara. 2012. [Lexicalization in signed languages: When is an idiom not an idiom?](#) In *Selected Papers from UK-CLA Meetings*, volume 1, pages 229–248. UK-CLA.
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. [Öffentliches DGS-Korpus: Annotationskonventionen / Public DGS Corpus: Annotation conventions](#). Project Note AP03-2018-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022. [The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).
- Thomas Krause and Amir Zeldes. 2016. [ANNIS3: A new architecture for generic corpus query and visualization](#). *Digital Scholarship in the Humanities*, 31(1):118–139.
- Gabriele Langer, Anke Müller, Felicitas Otte, and Sabrina Wähl. 2022. [Information types and use cases](#). Project Note AP10-2021-02, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Gabriele Langer, Anke Müller, Sabrina Wähl, Felicitas Otte, Lea Sepke, and Thomas Hanke. 2024. [Introducing the DW-DGS – the digital dictionary of DGS](#). In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy. European Language Resources Association (ELRA).
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [MediaPipe: A framework for perceiving and processing reality](#). In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*.
- Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and

- Arvid Schwarz. 2008. [DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).
- Marc Schulder and Thomas Hanke. 2020. [Open-Pose in the Public DGS Corpus](#). Project Note AP06-2019-01, DGS-Korpus project, IDGS, Hamburg University, Hamburg, Germany.
- Marc Schulder and Thomas Hanke. 2022. [How to be FAIR when you CARE: The DGS Corpus as a case study of open science resources for minority languages](#). In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 164–173, Marseille, France. European Language Resources Association (ELRA).
- Sehen statt Hören. 2006. Episode 1291: Die Haushaltshilfe. Broadcast by Bayrischer Rundfunk.
- Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. [Hand keypoint detection in single images using multiview bootstrapping](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4645–4653, Honolulu, Hawaii, USA.
- Erin Wilkinson, Ryan Lopic, and Lynn Hou. 2023. [Usage-based grammar: Multi-word expressions in American Sign Language](#). In Terry Janzen and Barbara Shaffer, editors, *Signed Language and Gesture Research in Cognitive Linguistics*, pages 357–388. De Gruyter Mouton, Berlin, Boston.

# Introducing the DW-DGS – The Digital Dictionary of DGS

Gabriele Langer , Anke Müller , Sabrina Wähl ,  
Felicitas Otte , Lea Sepke , Thomas Hanke 

Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany  
{gabriele.langer, sabrina.waehl, felicitas.otte, thomas.hanke}@uni-hamburg.de  
ankemueller.am@posteo.de, lea.sepke@outlook.de

## Abstract

This article describes the lexical resource DW-DGS – the first corpus-based digital dictionary of German Sign Language (DGS). Basic information is provided on dictionary type, context of compilation, sign representation in the product, metalanguage, dictionary content, information types displayed in entries, and dictionary structure. The article also provides an overview on data sources, methods, workflow procedures, and tools used in the lexicographic process. Challenges of making a corpus-based sign language dictionary and solutions developed for the DW-DGS are mentioned. The aim of this contribution is to provide an overview on the resource. It also serves as a starting point by referring to papers that describe the structures and procedures of the DW-DGS in more depth.

**Keywords:** sign language dictionary, lexical resource, German Sign Language (DGS), corpus-based lexicography

## 1. General Information

The full title of the online electronic dictionary of DGS described here is *Digitales Wörterbuch DGS (DW-DGS). Das korpusbasierte Wörterbuch DGS – Deutsch* [Digital Dictionary of DGS (DW-DGS) – German]. It is one of the products of the *DGS-Korpus project* (2009–2027).

The *DW-DGS* can be accessed at: <https://dw-dgs.de>. We refrained from including screenshots from the dictionary as figures and ask the reader to open the online dictionary for illustration. We recommend to look at entries 193, 366, 440 and 354 that cover most information types mentioned in this paper.

## 2. Dictionary Type

The *DW-DGS* is the first general corpus-based dictionary of German Sign Language (DGS). It is a descriptive dictionary produced in an academic context that focuses on the documentation of the general language of DGS. As a synchronic dictionary it targets at contemporary language – based on DGS as it was used at the time of data collection (2010–2012). The dictionary covers signs from all regions of Germany.

In the *DW-DGS*, established DGS signs are described from a primarily monolingual perspective on the basis of their uses in context as evidenced in the data of the *DGS Corpus*. The dictionary is corpus-based and largely, but not completely corpus-bound. Following the well-established corpus-based approach of modern lexicography, the results of corpus analyses for each lemma sign are summarised in the dictionary entries (cf. e.g. Atkins

and Rundell, 2008; Sinclair, 2003). The metalanguage used for description is German.

In addition to the description of DGS signs from a monolingual perspective the dictionary also provides some bilingual features. Senses of DGS signs listed and described in the entries are matched to German translational equivalents. This enables using the *DW-DGS* in the function of a bilingual dictionary DGS→German. The German index provides access to the DGS entries via German words thus fulfilling the function of a bilingual dictionary German→DGS. The *DW-DGS* can therefore be described as monolingual dictionary of DGS with additional bilingual features, or as a bilingualised monolingual dictionary (cf. Hannay, 2003; Svensén, 2009).

As far as the medium and conditions of publication and use are concerned, the *DW-DGS* is an electronic online dictionary that can be accessed freely and free of charge on the internet. It includes video clips of signs and signed example sentences.

The *DW-DGS* is made for a wide audience of users including the user groups of L1 DGS signers, L2 learners of DGS, DGS teachers, DGS interpreters, linguists, and the interested public.

For a discussion on the dictionary type, languages in the *DW-DGS*, and user groups cf. Langer et al. (2018b, 2022) and Müller et al. (2022).

## 3. Data Sources

Information provided in the *DW-DGS* takes into consideration data from three sources: The main source used is the *DGS Corpus*. It is supplemented to a small extent by data elicited via the *DGS-Feedback* and through *SignHunter*.

### 3.1. DGS Corpus

The *DGS Corpus* has been designed explicitly with the aim to provide a basis for the first corpus-based dictionary of DGS. Its current size is more than 680.000 tokens (Feb 2024).

So far the dictionary is not only based on corpus data but also largely corpus-bound. We do not include signs or senses that are not evidenced in the corpus data. For reasons of reliability, information on meaning and usage are based on analyses of fluent signing in context. Entry information are abstractions from corpus evidence. For corpus analyses all *DGS Corpus* data available of a lemma sign are used. That includes data published in the *Public DGS Corpus* but also lemmatised unpublished data. The *DGS Corpus* data is stored, annotated and worked with in iLex. iLex is the lexical database and annotation environment that is used in the project for annotation, data curation and analyses. In iLex the data is matched to a hierarchy of type and subtype entries.

For further information on corpus design, elicitation tasks, data collection and corpus curation cf. [Schulder et al. \(2021\)](#). For the concept of types and subtypes and the type structure in the iLex database cf. [Langer et al. \(2018a\)](#). [Langer et al. \(2016b\)](#), especially the poster, includes an example illustrating the different type levels and their use in the iLex database. Type levels as displayed in the *Public DGS Corpus* are explained in [Konrad et al. \(2022\)](#). For more information on iLex cf. [Hanke \(2002\)](#).

### 3.2. DGS-Feedback

Some usage data for a small set of signs have been collected online from signers via the so-called *DGS-Feedback*. Participants were presented sign-meaning combinations and asked whether they used or knew these signs for these meanings. These data are used in addition to complement, clarify or solidify the results from corpus data.

More information on how the data collected by the *DGS-Feedback* is used in compiling entries for the dictionary cf. [Wähl et al. \(2018\)](#). For a description of the *DGS-Feedback*, the design of the questionnaires and the question types cf. [Matthes et al. \(2014\)](#). For technical aspects of the *DGS-Feedback* system cf. [Berding and Hanke \(2015\)](#).

### 3.3. SignHunter

*SignHunter* is a tool that was created and used to collect additional data from participants at deaf events. Participants are presented isolated stimulus items and can choose items for which they want to contribute and record their signs. *SignHunter* was used merely for concepts that were consid-

ered unambiguous. So far signs for city names and signs for social media names have been collected via *SignHunter*.

For more information on the data collection tool *SignHunter* cf. [Hanke et al. \(2020\)](#).

## 4. Representation of Signs

There is no established, widely known writing system for DGS that could be used to represent DGS in the dictionary. As we do not expect the occasional user to learn a notation system just to be able to consult the dictionary, we decided against using notations. Glosses were not an option either: They bear the risk of interference by the gloss word and conflict with the idea of representing signs as entities of their own, spoken-language independent visual nature. Instead, signs are represented either by recorded videos or by small visual elements called micons. A micon is a thumbnail movie displaying the form of the lemma sign combined with a unique entry number for quick identification and reference. Hovering over the micon's thumbnail sets the micon in motion, clicking on it plays a larger video of the sign in the movie display area, clicking on the number below the thumbnail opens the corresponding entry.<sup>1</sup>

In the entries, micons are used as sign representations for information types given in DGS such as synonyms, antonyms, collocational patterns and multi-sign expressions where they also serve as implicit cross-references. Outside the entries, micons are used to represent lemma signs in access structures.

For more detail on the rationale of a gloss-free dictionary and the use and function of micons as a means of lemma sign representation cf. [Langer et al. \(2018b, 2022, 2019\)](#) and [Otte et al. \(2022\)](#).

## 5. Metalanguage

Written German is not only one of the target languages of the dictionary, but it is also used for the dictionary definitions, descriptions, comments and subject categories in the entry, as well as category headings and other elements used for orientation or navigation, such as menu options and buttons.

Front matter information is also provided in written German. While a signed version is not yet complete, users of DGS find related information in a set of tutorials explaining the *DW-DGS* in DGS.

---

<sup>1</sup>The term 'micon' is derived from 'moving icon' and was first coined by Russel Sasnett ([Brøndmo and Davenport, 1989](#)). In this original use, 'micon' referred to the small video playing in loop on its own. We have adapted the term for our purposes to include the ID number as well.



The tutorials in DGS can be found at <https://dgs-korpus.de/tutorials.html>. On the rationale for using German as metalanguage cf. Langer et al. (2022) and Müller et al. (2022).

## 6. Content of the DW-DGS

### 6.1. Signs

The dictionary describes established manual signs of DGS. Only simplex signs are treated as lemma signs and are given entry status. Multi-sign expressions aka multi-word expressions (MWE) are not treated as lemma signs in their own right. They are to some extent included and appear within entries at different places, either on the sense level, as information addressed to a sense, or in the run-on section of an entry. For the time being there are no entries for productive forms i.e. classifier signs or classifier handshapes, nor for non-manual elements.

Signs and senses listed in the *DW-DGS* are largely restricted to what is evidenced in the *DGS Corpus*. Sign variants, that is lexical and phonological variants, are included. Lemma selection is guided by frequency.

### 6.2. Information Types

The entries of the *DW-DGS* contain several different kinds of information. The following information types relate to the lemma sign as a whole:

1. *Form*: information on form and form variants provided as studio recordings;
2. *Kommentar*: comments on aspects of form, usage and other additional information on the sign;
3. *Beleglage*: rough indication of frequency of the sign in the *DGS Corpus*;
4. *Grammatik*: grammatical label or comment;
5. *Regional*: comment on regional distribution, including distributional maps;
6. *Bedeutung*: information on meaning and use: list of senses in the form of signposts <sup>2</sup>;
7. *Zusammensetzungen*: compound-like constructions containing the sign;
8. *Verwandt/Formgleich/Formähnlich*: cross references to related signs and signs of the same form or a similar form, and

<sup>2</sup>Within the list of senses some MWE are listed under the categories phrase (*Phrase*) and multi-sign name (*Mehrteiliger Name*).

9. *Konkordanz*: concordance view of tokens of the sign in the *Public DGS Corpus*.

The following information types relate to a particular sense in the senses' section:

10. rough indication on the meaning of the sense (*Signpost*);
11. *Form* (only provided for phrases and multi-sign names);
12. *Mundbild*: selection of typical mouthings or information on mouth gesture used with the sign and a studio recording of the sign with a typical mouthing or mouth gesture;
13. *Erklärung*: explanation of the sense, the so-called dictionary definition;
14. *Deutsch*: German translational equivalents, sometimes with disambiguation information or diasystematic label;
15. *Anmerkung*: additional information on usage;
16. *Grammatik*: grammatical information specific to the sense;
17. *Beispiele*: authentic examples illustrating the sense, each with a clip of the original *DGS Corpus* recording, a German translation and a short context, and with direct links to its original location in the two portals of the *Public DGS Corpus* (*MY DGS* and *MY DGS – annotated*);
18. *Bedeutungsgleich*: synonym and near-synonym signs, sometimes with a clickable thumbnail map that displays the regional distribution of a set of coexisting lexical regional variants;
19. *Entgegengesetzt*: antonym signs of opposite or complementary meaning sometimes with a clickable thumbnail map that displays the regional distribution of a set of coexisting lexical regional variants;
20. *Häufige Kombinationen*: collocational patterns and semantic preference patterns;
21. *Zusammensetzungen*: compound-like constructions that can be related to this particular sense;
22. *Regional*: comment on the regional distribution of this sense or a group of senses with a link to a corresponding distribution map;
23. *Sachgruppen*: subject areas that this sense is assigned to.

Not all types of information are given in each entry or for each sense. Information is provided only when relevant and available.

Langer et al. (2022) provides further details on the information types mentioned here.

### 6.3. Types of Entries

Entries in the *DW-DGS* differ with respect to their analytic and descriptive depth. This is partly due to the varying amount and quality of data available in the corpus for each lemma sign and partly due to issues of time and resources.

While lexicographers explored what could be done with corpus data at hand for a larger number of entries, it was not possible to invest the same amount of time and labour in the preparation of all entries in the same way. As a consequence the team opted for a mixture of more and less elaborated entries. The dictionary contains elaborated entries and shorter entries with less fine-grained sense distinctions and less information. Also, there are entries completely edited by lexicographers and entries that have been partly edited starting with automatically compiled data.

A third kind of entries in the *DW-DGS* are automatically compiled entries. For these entries only minimal editing steps such as lemma establishment were done manually. Information provided in automatically compiled entries includes senses inferred from cross-references originating at manually edited entries as well as rough meaning indications, i.e. German equivalents, already prepared in iLex for subtypes of the *DGS Corpus* types list. Automatically compiled entries have not yet received a full lexicographic treatment. Such entries are not included in the *DGS* index but can be accessed from cross-references addressing them in edited entries and through their listing in the German index. Automatically compiled entries can be identified by their micon appearance (red number on white background as opposed to the white number on red background shown for edited and partly edited entries) and by the heading (*Automatisch generierter Vorabbeitrag*) at the top of the entry page.

For more information on different entry types in the *DW-DGS* cf. Wähl et al. (2022).

## 7. Dictionary Structure

### 7.1. Navigation: Menu Bar

The menu bar at the top of each *DW-DGS* web page enables the user to choose which part of the dictionary they want to visit. The default page is the body of entries (option: *DGS*) which is displayed by default when opening the dictionary URL.

### 7.2. Front Matter (*Intro*)

The front matter of the *DW-DGS* contains an introduction including information on dictionary use (user's guide), background information on the data used, the lexicographic process, maps, and relevant object language information for *DGS*.

### 7.3. Back Matter (*Karten*)

In the back matter, the users find a number of maps including geographical distribution maps of coexisting regional signs belonging to specific semantic sets, such as signs for the days of the week or colors, and interactive geographical maps with city and country name signs.

### 7.4. Access Structures

While in print dictionaries there is one primary sort key determining the order of entries in the main part of the dictionary and several indexes on secondary sort keys, the *DW-DGS*, like many electronic dictionaries, has individual pages for each of the entries and several indexes providing access to the individual pages.

For more information on the access structures available in the *DW-DGS* cf. Langer et al. (2022).

#### 7.4.1. Macrostructure (*DGS*)

The main and most important index of the *DW-DGS* shows the body of entries. Each entry is represented by a micon. The macrostructure consists of a table of all micons. The user can choose between several options of ordering the entries represented by micons: by entry number (*Nummer*), which is also the default, by handshape (*Handform*), by number of hands (*Händigkeit*), or by place of articulation (*Lokation*). The secondary sort key is the height of the place of articulation from high to low. Where the variants of an entry differ with respect to their values for the current sort keys, the variants are shown separately, i.e. the micons then represent individual variants and are thus marked with a .1, .2, etc. appended to the entry number. This ordering allows for a very rough search by form.

#### 7.4.2. German Index (*Deutsch*)

The German index is an additional access way to the information provided in the sign entries and supports searches for signs through German words. It consists basically of a table with a listing of German words in the first column. In many cases the German word is disambiguated by a context in the second column. Micons represent the corresponding sign senses in the third column. In some cases of high regional lexical variation an additional thumbnail map is displayed in the third column as a cross

reference to the cluster map visualising regional distribution of coexisting lexical variants.

The German index is generated from the German translational equivalents provided in the entries.

The German words in the German index do not receive the same in-depth attention and treatment as lemma signs of their own right as they would in a fully bilingual dictionary. No missing words or word senses are added.

The German index provides dictionary-external links to corresponding entries in the corpus-based German Dictionary *DWDS* ([Berlin-brandenburgische Akademie der Wissenschaften](#)) where additional information on the German words can be looked up quickly if desired. This compensates somewhat for the scarceness of information provided for German equivalents.

### 7.4.3. Subject Area Index (*Sachgruppen*)

In the entries, each sense is matched to up to three subject areas in which the the sign-sense combination is then listed. The subject area index is a topic-specific way to access the signs contained the *DW-DGS*. In a table it lists subject areas together with the signposts and micons of the senses allocated to them.

### 7.4.4. Graph (*Graph*)

The graph is a visual structure that provides a non-text-based access way to the dictionary. Entries are depicted as dots. A clickable micon appears when the cursor hovers over a dot. The dots are connected by color-coded lines that represent different relation categories between entries, that is synonyms (*Bedeutungsgleich*), antonyms (*Entgegengesetzt*), collocations (*Häufige Kombinationen*), compoundlike constructions (*Zusammensetzungen*), parts of MWE (*Bestandteile*), signs having the same form (*Formgleich*), signs having a similar form (*Formähnlich*), and related signs (*Verwandt*). The user can modify the graph to show only certain kinds of connections by unclicking all other checkboxes. The graph is a tool for playful exploration of the dictionary.

For more information on visual access to the dictionary by the graph cf. [Langer et al. \(2022\)](#), a short description can be found in [Müller et al. \(2022\)](#).

### 7.5. Microstructure: Entries

Each entry has its own web page with a unique entry number for identification at the top, a video display area, and a table containing the entry information. In the first column of the table the category labels of the information types are given while the second column contains the information provided.

The head section shows information addressed to the whole sign. It is followed by the list of senses. At the bottom run-on information such as MWE and form-related cross references to other signs are given.

The middle part contains the list of senses. In its collapsed state it is presented as a list of signposts that hint on the senses' meanings. Each sense row is numbered and can be expanded to reveal all information addressed to the sense. When expanded, category labels for the information types addressed to the sense are displayed in the second column while the corresponding information is provided in an additional column to the right. Information given in DGS is either displayed as micons or can be viewed as movie in the video display area by clicking on the button with the play-symbol.

### 7.6. Mediostructure: Cross Referencing

In the *DW-DGS* all cross-references to lemma signs, variant forms, or senses are realized as micons. The thumbnail micon figure represents the lemma sign in the form of either the first variant as default or, when relevant, a different variant of the lemma sign. The specific address is expressed by the number of the micon: Cross-referenced lemma signs (i.e. whole entries) appear with the entry number only (e.g. 144), micons for cross-referenced variants with entry number followed by a point and the number of the variant form (e.g. 144.2), cross-referenced senses with entry number followed by a hash and the sense number(s) (e.g. 144#2). Micons function as implicit cross references as there is no special reference marker. Entry-internal and entry-external cross references are not distinguished visually.

Hyperlinks to entries in the German index or in the subject area index use the written German words as labels set in blue text color.

The dictionary contains two kinds of maps, one showing the regional distribution of a single lemma sign and the second showing several coexisting lexical, mostly regional variants in contrast to each other (cluster maps). For the first kind, *Karte* (map) hyperlinks lead to extra map pages. Cross references to cluster maps are realised by clickable thumbnails of the map.

## 8. Links to other Resources

The *DW-DGS* is an online resource that makes use of the possibility to include cross references that directly link to resources outside the dictionary. For more detail on linking to and from the *DW-DGS* cf. [Müller et al. \(2020\)](#).

## 8.1. Linking to the Public DGS Corpus

At two different places in the *DW-DGS* external links to the *Public DGS Corpus* are provided.

Example sentences link via buttons below the video display area to their location in the two portals of the *Public DGS Corpus*. These buttons are not shown when examples are taken from unpublished parts of the *DGS Corpus*.

At the very bottom of each entry the button *Konkordanz* opens a page with a concordance view of all tokens in the annotated Public DGS Corpus (*MY DGS – annotated*) that are realisations of the lemma sign of the respective *DW-DGS* entry. From the concordance view one can jump to the corresponding types list entry of *MY DGS – annotated* by clicking on a gloss or into the transcript by clicking on the transcript's name in the upper left side of a concordance line.

## 8.2. Linking to the DWDS

In the German index, links to the German online dictionary called *DWDS* are provided whenever a match could be found for a listed German word or word sense. The links are realised in form of clickable blue and white *DWDS* logos.

# 9. Method and Workflow

For the lexicographic process from data analysis to the finished dictionary entry, we adapted the lexicographic principles and steps of corpus-based lexicography as described in [Atkins and Rundell \(2008\)](#) for sign language lexicography.

For a short description of the steps in the lexicographic process cf. [Langer \(2021\)](#) and [Langer et al. \(2018a\)](#).

## 9.1. Dictionary Writing System

We use a FileMaker database as our dictionary writing system (DWS). Filemaker is a low-code program in which the user is able to configure and adjust the user interface without programmer's support. Some information from iLex such as type glosses and HamNoSys notations is imported for direct display into the DWS. Other entities are entered with only their iLex ids for reference, e.g. types used for cross references and tags needed for authentic example management. These entities as well as SQL queries can be opened directly in iLex via scripts stored in the FileMaker database. In the DWS, types and subtypes are grouped for lemma establishment, pre-lexicographic information is stored, and entry information is prepared and edited for publication.

## 9.2. Lemma Selection

The first step of the lexicographic process is lemma selection. For the *DW-DGS* this is basically driven by frequency. The general threshold for inclusion of a corpus type candidate into the dictionary is that it contains at least one subtype with at least 25 tokens from a number of different signers. In certain cases we work with less than 25 tokens, for example when we are dealing with lexical variants that are used only in certain geographic areas and are part of a group of several coexisting regional signs for one concept, or when the sign in question is part of a semantic set that would be missing one element just because of low token numbers. Tokens of phonological variants may be added up to meet the threshold while so-called non-tokens should be excluded.

Lemma selection is described in more detail in [Wähl et al. \(2022\)](#). A description and discussion of non-tokens can be found in [Langer et al. \(2016a\)](#).

## 9.3. Establishment of Lemma Signs

For each selected lemma sign candidate it has to be determined which subset of tokens, i.e. which types and subtypes are apt to constitute the data to be described in this sign's dictionary entry. During this step also other variants, related or similar types have to be checked for possible inclusion. Inspired by [Svensén \(2009, p. 94\)](#), we call this establishment of lemma signs to distinguish it from lemmatisation in annotation.

For DGS the establishment of lemma signs is much less straightforward than for a well researched spoken language with a long codified written tradition. In DGS we find a high variation in form, iconic modifications of sign forms, and a somewhat flexible combination of signs with mouthings that contribute to the semantics of the signs. Often the lexicographers are confronted with a large number of similar signs with only partly overlapping meanings. This makes lemma establishment in DGS a rather challenging task. The process requires a number of different aspects, principles and criteria that have to be taken in consideration and weighed against each other. While in principle lemma selection and the establishment of lemma signs are two separate steps, in practice they are mutually dependent and thus done at the same time.

The lemma establishment rules and principles used for the *DW-DGS* with illustrating examples can be found in [Langer et al. \(2020\)](#). [Hanke et al. \(2023\)](#) describes an example where the regional distribution of subtypes is considered in the decision making for the establishment of lemma signs. For sign languages issues of lemma sign establishment have also been described and discussed by [Johnston and Schembri \(1999\)](#); [Kristoffersen and](#)



Troelsgård (2010); Fenlon et al. (2015).

## 9.4. Compiling Entries

Compiling entries is a complex task in which analysing available data, abstracting, summarising and describing the results while preparing the entries for publication goes back and forth. Corpus data analyses during this task include a look at form variation and sign forms in fluent signing, regional distribution, distribution across age groups, and frequent neighbours (collocations). The citation form and variant forms to be included are determined. The central and most time-intensive task is Word Sense Discrimination (WSD). Once the senses of a lemma sign have been determined and described in the DWS all other information addressed to a sense can be entered and prepared for publication.

The various steps during compiling entries of the *DW-DGS* are described in Langer (2021); Langer et al. (2018a).

### 9.4.1. Word Sense Discrimination (WSD)

WSD encompasses identifying meanings and uses of a lemma sign and describing them as a list of sign senses. For this task a considerable number of tokens are viewed in context, that is, the recorded movies are viewed, alongside with annotations and translations. Lists of frequent neighbours in the corpus can help to identify different uses. Results of recurring similar uses are summarised and entered as senses in the DWS database. Each sense is described by a German explanation, the so-called dictionary definition. Corpus tokens contributing to the evidence for a sense are tagged in iLex and suitable candidates for dictionary examples are referenced in the DWS for further use. In a second step the proto-senses are reviewed, lumped when necessary, and marked for production or exclusion.

### 9.4.2. Editing the Entry

Usage examples are selected from the candidates to illustrate meanings and typical uses of a sign in context. They are prepared for publication and receive a short context and the translation is adapted for use in the dictionary. Synonymous or antonymous signs are included in the description of a sense if available in order to provide language-internal hints on the sign's sense. Further information that is part of a sign sense's section is prepared and entered into the DWS, including typical mouthings, sign combinations such as collocations and semantic preferences, German translational equivalents, subject areas, and information on usage and distribution across regions and age

groups. Maps showing regional distribution are prepared and marked for publication when subtypes display a noticeable regional distribution. When necessary, disambiguating contexts are entered for translational equivalents. Equivalents that are less useful in a reversal of the search direction from *DGS*→German to German→*DGS* are marked for exclusion from the German index. Each sense also is provided a signpost for the meanings overview in collapsed entries.

For the extraction of a frequent neighbour lists from the *DGS* corpus and its use in analyses and inclusion in the *DW-DGS* cf. Langer (2021); Langer et al. (2018a); Langer and Schulder (2020). On the selection and preparation of authentic examples from the corpus cf. Langer et al. (2018b). More information on the use of maps for exploration in analyses and an example analysis of regional clusters is found in Hanke et al. (2023).

## 9.5. Production

### 9.5.1. Studio Recordings

Representing *DGS* in the dictionary requires studio recordings of the respective signs. The videos are needed to produce the still and the animated figure of the micon as well as to be played in the video display area to show the sign's form in isolation. At the level of the lemma sign, signs are recorded without mouthing to serve as information on variant forms. For each sense, an individual recording is made showing the sign with a typical mouthing or mouth gesture.

A list of required movies is generated from the DWS, matched against the list of movies already available in iLex from previous recordings, and provided as a script list for the studio recording. The required signs are performed by deaf models in our video studio. In total, seven different cameras from four different angles are used. In post-production movies are converted and integrated into iLex. Then the video material is annotated by student assistants and checked by deaf team members for whether the signs were correctly executed by the signing model. After that student assistants choose one of the frames of the video as the thumbnail for the micons.

### 9.5.2. Production of the Dictionary

For the production of an updated dictionary version relevant data for the prepared entries are exported from the DWS and converted into a json file. The production scripts generate maps, video clips of studio recordings and example sentence videos as well as the micons. Production data from the json file and data from iLex are combined to generate the html pages of the dictionary for the manually edited

entries as well as for the automatically compiled entries.

On technical aspects of generating maps from iLex data cf. [Hanke \(2018\)](#).

## 10. Highlights of the DW-DGS

The *DW-DGS* contains several interesting and new information types with regard to sign language dictionaries. Here only few highlights can be mentioned briefly.

One special trait of the *DW-DGS* are the authentic examples taken from the original recorded data. They illustrate and reinforce the more general and abstract sense descriptions and contribute an element of liveliness. In addition, the example movies grant visibility of the language community through its contributing members.

Through the corpus it has become possible to use statistics to identify frequent neighbours in DGS. To our knowledge it is the first time that collocation patterns were used for WSD of a sign language and are included as information in a sign language dictionary.

Maps showing the regional distribution for individual signs or several coexisting regional signs (cluster maps) are very attractive for many deaf users. These maps are directly generated from corpus data and participant metadata and would not be possible without data from the *DGS Corpus*.

The last information type provided in the *DW-DGS* that we want to highlight here are the cross references to sign synonyms, near synonyms and antonyms of a sense. They serve as monolingual explanatory elements. A robust and proper distinction and description of sign senses is a good basis and in our view almost a prerequisite for discovering and displaying information on synonym and antonym relations between signs.

## 11. The DW-DGS in Numbers

The *DW-DGS* contains 1876 entries: 802 fully edited, 272 partly edited entries and 802 automatically compiled entries (cf. section 6.3. *Types of Entries* for more details on entry types). The 1074 edited entries contain 2436 senses with 3416 authentic examples and 6377 German equivalents. For 581 senses sense-related distribution maps are provided. The dictionary includes 50 distribution maps for sets of coexisting regional lexical variants. Additional 1290 cross-references between synonyms and near-synonyms and 726 cross-references between antonyms are provided between senses. The edited entries include 416 collocational patterns and 141 compound-like patterns. There are 2679 manually selected cross-

references to signs of similar or same form between entries. (Numbers date from end of March 2024.)

## 12. SL-specific Challenges

There are many challenges specific to corpus-based sign language lexicography. A high variation of sign forms with many partly overlapping meanings in a corpus of still limited size make lemma sign establishment complicated.

The lack of a written direct representation of the signing in the corpus along with only limited tools for corpus annotation and analysis makes working with signed corpus data a very time-consuming task as lexicographers cannot skim-read through samples but have to resort to watching the original video data one by one and sometimes several times to compare.

For the design and structure of a dictionary, the lack of a writing system and orthography for signed text results in the issues of sign representation, entry ordering (macrostructure), search for sign form, and the choice of metalanguage in the dictionary.

For a short discussion of some of these challenges cf. [Müller et al. \(2022\)](#).

## 13. Outlook

Corpus annotation is on-going. This enables us to expand existing entries as well as to create new ones. The *DW-DGS* is updated several times a year.

For searching a specific sign by its form, the *DW-DGS* currently offers different sort orders for the body of entries. There, the sign looked for then needs to be identified by browsing through the corresponding subsection. This becomes tedious if such a subsection contains too many items or the user is unsure about the location used in the sign (the secondary order criterion). Filtering facilities as implemented e.g. by the *ODT* ([Centre for Sign Language, 2008–ongoing](#)) and the *GaLex* ([Konrad et al., 2010](#)) might be of some help here, but this approach needs to be tailored to the size of entries in order to provide result sets small enough that browsing after filtering still appears natural to the user. So such a functionality can only be implemented now that the dictionary growth is expected to be much slower than in the past years.

## 14. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies'

Programme is coordinated by the Union of the Academies of Sciences and Humanities.

## 15. Bibliographical References

- B. T. Sue Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford; New York.
- Sven Berding and Thomas Hanke. 2015. [Documentation of the Feedback-System and its integration into iLex](#). Project Note AP04-2015-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Hans Peter Brøndmo and Glorianna Davenport. 1989. Creating and viewing the elastic charles – A hypermedia journal. In *Hypertext: State of the Art. Papers presented at the Hypertext 2 conference, York, 1989*, pages 43–51. Intellect Ltd., Oxford.
- Jordan Fenlon, Kearsy Cormier, and Adam Schembri. 2015. [Building BSL SignBank: The Lemma Dilemma Revisited](#). *International Journal of Lexicography*, 28(2):169–206.
- Thomas Hanke. 2002. [iLex - A Tool for Sign Language Lexicography and Corpus Analysis](#). In *3rd International Conference on Language Resources and Evaluation*, pages 923–926, Las Palmas, Canary Islands, Spain. European Language Resources Association (ELRA).
- Thomas Hanke. 2018. [Using d3js to visualise data in iLex](#). Project Note AP04-2016-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Thomas Hanke, Elena Jahn, Sabrina Wähl, Oliver Böse, and Lutz König. 2020. [SignHunter – A Sign Elicitation Tool Suitable for Deaf Events](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 83–88, Marseille, France. European Language Resources Association (ELRA).
- Thomas Hanke, Reiner Konrad, and Gabriele Langer. 2023. [Exploring regional variation in the DGS Corpus](#). In Ella Wehrmeyer, editor, *Advances in Sign Language Corpus Linguistics*, number 108 in Studies in Corpus Linguistics, pages 192–218. Benjamins, Amsterdam.
- Mike Hannay. 2003. [3.1 Types of bilingual dictionaries](#). In Piet van Sterkenburg, editor, *Terminology and Lexicography Research and Practice*, volume 6, pages 145–153. Benjamins, Amsterdam.
- Trevor Johnston and Adam C. Schembri. 1999. [On Defining Lexeme in a Signed Language](#). *Sign Language & Linguistics*, 2(2):115–185.
- Reiner Konrad, Thomas Hanke, Gabriele Langer, Susanne König, Lutz König, Rie Nishio, and Anja Regen. 2022. [Public DGS Corpus: Annotation conventions](#). Project Note AP03-2018-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Jette Hedegaard Kristoffersen and Thomas Troelsgård. 2010. Making a Dictionary without words: Lemmatization problems in a sign language dictionary. In Sylviane Granger, editor, *2010 eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*, number 7 in Cahiers Du Cental, pages 165–172. Presses Univ. de Louvain, Louvain.
- Gabriele Langer. 2021. [Vorgehen bei der Analyse für die Artikelschreibung \(Wörterbuch\)](#). Arbeitspapier AP10-2016-02, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Gabriele Langer, Thomas Hanke, Reiner Konrad, and Susanne König. 2016a. [“Non-tokens”: When Tokens Should not Count as Evidence of Sign Use](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 137–142, Portorož, Slovenia. European Language Resources Association (ELRA).
- Gabriele Langer, Anke Müller, Felicitas Otte, and Sabrina Wähl. 2022. [Information Types and Use Cases](#). Project Note AP10-2021-02, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Gabriele Langer, Anke Müller, and Sabrina Wähl. 2018a. [Queries and Views in iLex to Support Corpus-based Lexicographic Work on German Sign Language \(DGS\)](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 107–114, Miyazaki, Japan. European Language Resources Association (ELRA).
- Gabriele Langer, Anke Müller, Sabrina Wähl, and Julian Bleicken. 2018b. [Authentic Examples in a Corpus-Based Sign Language Dictionary – Why and How](#). In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2018)*, pages



- 483–497, Ljubljana, Slovenia. Ljubljana University Press.
- Gabriele Langer, Anke Müller, Sabrina Wähl, and Thomas Hanke. 2019. [The DGS-Korpus approach to including frequent sign combinations in a corpus-based electronic sign language dictionary](#).
- Gabriele Langer, Anke Müller, Sabrina Wähl, Susanne König, Thomas Hanke, and Reiner Konrad. 2020. [Lemmatisierungsregeln \(Vorläufige Wörterbuch-Einträge\)](#). Arbeitspapier AP10-2016-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Gabriele Langer and Marc Schulder. 2020. [Collocations in Sign Language Lexicography: Towards Semantic Abstractions for Word Sense Discrimination](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 127–134, Marseille, France. European Language Resources Association (ELRA).
- Gabriele Langer, Thomas Troelsgård, Jette Kristoffersen, Reiner Konrad, Thomas Hanke, and Susanne König. 2016b. [Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 143–152, Portorož, Slovenia. European Language Resources Association (ELRA).
- Silke Matthes, Thomas Hanke, Susanne König, and Gabriele Langer. 2014. [Entwicklung des DGS Feedback-Systems](#). Arbeitspapier AP07-2014-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Anke Müller, Thomas Hanke, Reiner Konrad, Gabriele Langer, and Sabrina Wähl. 2020. [From Dictionary to Corpus and Back Again – Linking Heterogeneous Language Resources for DGS](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 157–164, Marseille, France. European Language Resources Association (ELRA).
- Anke Müller, Gabriele Langer, Felicitas Otte, and Sabrina Wähl. 2022. [Creating a dictionary of a signed minority language. A bilingualized monolingual dictionary of German Sign Language](#). In *Proceedings of the XX EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2022)*, pages 635–648, Mannheim, Germany. IDS-Verlag.
- Felicitas Otte, Anke Müller, Gabriele Langer, Sabrina Wähl, and Thomas Hanke. 2022. [Sign representation in the DW-DGS](#). Project Note AP11-2021-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- Marc Schulder, Dolly Blanck, Thomas Hanke, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Lutz König, Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio, and Christian Rathmann. 2021. [Data statement for the Public DGS Corpus](#). Project Note AP06-2020-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.
- John Sinclair. 2003. Corpora for lexicography. In Piet van Sterkenburg, editor, *A Practical Guide to Lexicography*. Benjamins, Amsterdam.
- Bo Svensén. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press, New York.
- Sabrina Wähl, Gabriele Langer, and Anke Müller. 2018. [Hand in Hand - Using Data from an Online Survey System to Support Lexicographic Work](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 199–206, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sabrina Wähl, Gabriele Langer, Felicitas Otte, Anke Müller, Thomas Hanke, and Reiner Konrad. 2022. [Lemma Selection and Entry Types](#). Project Note AP10-2021-01, Universität Hamburg, DGS-Korpus project, IDGS, Universität Hamburg.

## 16. Language Resource References

- Berlin-brandenburgische Akademie der Wissenschaften. [DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart](#).
- Centre for Sign Language. 2008–ongoing. [Ordbog over Dansk Tegnsprog. \[Dictionary of Danish Sign Language.\]](#)
- Reiner Konrad, Gabriele Langer, Susanne König, Thomas Hanke, and Christian Rathmann. 2010. [Fachgebärdlexikon Gärtnerei und Landschaftsbau](#).



# Annotation of LSF subtitled videos without a pre-existing dictionary

Julie Lascar, Michèle Gouiffès, Annelies Braffort, Claire Danet

Université Paris-Saclay, CNRS, LISN

Campus Universitaire Bat 507, rue du Belvédère, 91405 Orsay, France

julie.lascar, michele.gouiffes,annelies.braffort@lisn.upsaclay.fr, claire.danet@gmail.com

## Abstract

This paper proposes a method for the automatic annotation of lexical units in LSF videos, using a subtitled corpus without annotation. This method, based on machine learning and involving linguists for added precision and reliability, comprises several stages. The first consists of building a bilingual lexicon (including potential variants of a given lexical unit) in a weakly supervised manner. The resulting lexicon is then refined and cleaned by LSF experts. This data serves next to train a supervised classifier for automatic annotation of lexical units on the Mediapi-*RGB* corpus. Our Pytorch [implementation](#) is publicly available.

**Keywords:** French Sign Language, bilingual lexicon, sign spotting, automatic annotation

## 1. Introduction

Sign languages (SL) are natural languages used in Deaf communities. Their visuo-gestural nature allows information to be conveyed simultaneously using multiple articulators (hands, arms, body and facial components). SL content, where iconicity plays a central role, is spatially organised. The analysis of SL videos for annotation, recognition or translation requires the design of appropriate computer vision and natural language processing methods. A large amount of data is also needed, for instance videos with annotations, translations or subtitles. However, this kind of data is still scarce, particularly for French Sign Language (LSF).

Our study aims to devise a method for annotating videos with lexical signs with utmost precision while simultaneously reducing the manual annotation time required by experts.

After a short review on the related works (section 2), the paper describes a three-stages approach for automatic annotating LSF videos subtitled in French. The first stage (section 3) consists in a weakly supervised segmentation of specific signs in the videos, without use of any isolated example. The quality of the outputs is next assessed by LSF experts (section 4). Then, a supervised classifier (section 5) is trained using the previous annotations. In Section 6, experiments are conducted to investigate the impact of expert analysis on supervised classification and the scalability of our model.

## 2. Related works

The automatic annotation of lexical units in a SL video consists in determining the presence of such units and their temporal localization. We are therefore interested in *sign-spotting* approaches, which highly rely on *video encoding* methods. Regarding

LSF, there is unfortunately a scarcity of data for effective automatic processing.

**Sign spotting in continuous videos.** Sign spotting consists in localizing a sign temporally in a continuous video given a query. This is generally done by computing similarities between an example of the query sign and the video, and finding local maxima. While first works (Yang et al., 2009; Buehler et al., 2009) relied on similarities computed from hand-crafted features and involved limited dictionaries, more recent methods use learned classifiers, as in Jiang et al. (2021) where a transformer architecture is used. When available, subtitles can be used for a weak supervision as in Momeni et al. (2020), where multiple instance learning is leveraged. In Albanie et al. (2020), multiple modalities are used in the sign spotting, such as “mouthing”. These approaches rely on a dictionary of isolated signs, which is not available for all SL.

In Momeni et al. (2022), several methods are proposed to increase the density of annotations on continuous signing data. For instance, they localize unknown signs (not present in a lexicon), by selecting keywords in subtitles and finding the corresponding signs within continuous signing data. Our work enriches this technique to precisely locate the beginning and end of a sign.

**Video encoding.** The choice of the video encoding has a large impact on sign spotting performances. Most SL recognition models are inspired from the action recognition domain. First of all, a large number of works use pose-based representations to encode videos (Belissen et al., 2020b; Ouakrim et al., 2023); it has advantages for SL, in particular invariance with respect to the setting and the appearance of the signer, to keep only the gesture information and, to a certain extent,

facial expressions. However, recent studies have obtained some very interesting results, by using pre-trained models designed for action recognition in videos based on RGB images. Examples of such models include I3D (Carreira and Zisserman, 2017) and the more recent Video Swin Transformer (Liu et al., 2022). Fine-tuning these models specifically for sign recognition tasks yields impressive results, as demonstrated in tasks like fingerspelling recognition (Prajwal et al., 2022), sign spotting (Momeni et al., 2022) or translation (Li et al., 2020).

**LSF resources.** These various methods require the use of large quantities of data. However, many SL are under-resourced, such as for LSF (Kopf et al., 2022). Note however the 8h dialogues dataset DictaSign (Belissen et al., 2020) which is partly annotated by linguists and useful to recognize signs in context whether they are lexical (Ouakrim et al., 2023) or non-lexical (Belissen et al., 2020a,b). Recently, the corpus Mediapi-RGB has been released (Bull et al., 2024) comprises 86 hours of videos in LSF produced by deaf journalists or presenters from the bilingual online media Média’Pi<sup>1</sup>, with French subtitles produced by Deaf translators (Ouakrim et al., 2024). During a post-production phase, the videos are subtitled by professional translators. These translations are manually aligned with the corresponding SL video content. Mediapi-RGB is therefore, by construction, a perfectly aligned bilingual corpus. Our annotation system is built upon this dataset.

### 3. Step 1: Weakly supervised annotation

The lack of a freely available bilingual LSF/French dictionary led us to build our own one, by using the bilingual Mediapi-RGB corpus.

#### 3.1. French vocabulary choice

The first step implies to draw up a list of French words for which the corresponding signs are searched in the videos. The initial list was established from the subtitles by selecting lexical terms belonging to defined categories: days of the week, months, cities, countries, sports, vocabulary linked to current events (mask, unemployment, yellow waistcoats, film, etc.). These words were selected because they appear frequently in the dataset and are supposed to have stable meaning depending on the context.

For each word of this list, all video clips representing its LSF equivalent have to be segmented automatically and precisely, i.e. the full sign has to

be detected, with less transitions as possible. The main difficulty is the lack of isolated examples of the signs to be detected, since the videos are subtitled but not annotated. The method used for this task is outlined in the next section (3.2).

#### 3.2. Sign Spotting method

The technique described in Momeni et al. (2022) is used with different settings in order to fit to our dataset and our own objectives. Let us describe its principle on an example shown in Figure 1(a).

In this example, the objective is to capture the visual representation of “rugby” in a reference video. A similarity matrix (with values ranging in  $[0, 1]$ ) is computed between this reference video and  $N$  other positive examples, which are videos with subtitles containing the word “rugby”. For each of the  $N$  matrices, the maximum similarity value of each row is kept, leading to  $N$  vectors that are next aggregated using a voted scheme (threshold set at 0.6). This results in a vector  $L^+$ , which shows areas of significantly high similarity between the reference video and the positive examples. In these areas of high similarity, it is very likely to find the sign corresponding to “rugby”, but it may contain other frames belonging to transitions, or even signs that often appear in the same context. To avoid capturing these frames, the process is repeated using  $N'$  negative examples, i.e. videos for which the subtitle does not contain the word “rugby”. It yields a vector  $L^-$  which is useful to locate these non-positive frames. Finally, the vector  $L = L^+ - L^-$  improves the localization of the visual representation(s) corresponding to the word “rugby”. Unlike Momeni et al. (2022), our video clips have various sizes: a video clip is made for any consecutive sequence of at least 3 frames for which  $L$  is above a fixed threshold (set at 0.5). The maximum number of positive videos  $N$  is set to 100, and  $N'$  is set to  $3 \times N$ . Since the effectiveness of this method heavily relies on the way videos are encoded, the next section discusses the choice of video encoding.

#### 3.3. Videos encoding

To optimise the performance of this encoding step, three methods are compared. Figure 1(b) shows three similarity matrices computed between two videos which are supposed to contain the sign corresponding to the word “village”. The same couple of videos is used in each case, with the following subtitles on the vertical axis “A l’arrière, c’est-à-dire dans les *villages*, comme tous les hommes sont partis combattre,” (“In the rear, i.e. in the villages, as all the men had left to fight,”), and on the horizontal axis: “126 villes ou *villages* ont été placés en état de catastrophe naturelle.” (“126 towns and villages have been declared natural disasters.”).

<sup>1</sup><https://www.media-pi.fr/>

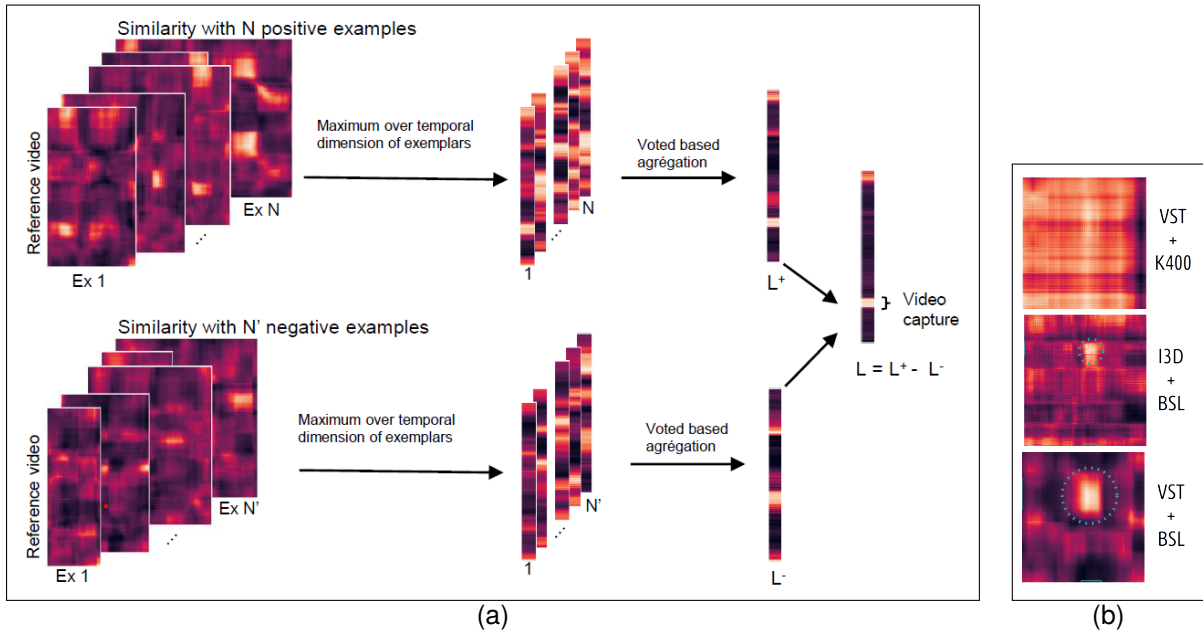


Figure 1: (a) Detecting unknown signs by spotting through exemplars. (b) In this example, we calculated the frame-by-frame cosine similarity between two videos with the word “village” in the subtitle, which had previously been encoded with 3 different backbones. The lightest area is the one with the greatest similarity, allowing us to locate the image sequences that visually represent the word “village”.

Each matrix relies on a different video encoder: at the top, a Video Swin Transformer (VST) trained for action recognition with Kinetic 400 (Liu et al., 2022); in the middle, an I3D model trained for sign recognition with BSL data (Renz et al., 2021); at the bottom, a Video Swin Transformer also trained on sign recognition with BSL data (Prajwal et al., 2022; Bull, 2023). The latter is selected for our study since it clearly provides the most discriminant similarity.

### 3.4. Refinement of the method

After this stage, various errors may occur and need to be thoroughly investigated and eliminated, as automatically as possible.

#### Dealing with the variability of form or meaning.

Some signs may vary in form depending on the signer. For example, some signs representing the months can be made with one hand or with both hands, depending on the signer. Others can also be completely different in form. In these cases, the method fails in finding similarities. To overcome this problem, the videos are automatically clustered by signer<sup>2</sup> when the number of positives examples is high enough (superior to 20), before applying the similarity search.

In addition, the method could fail due to the polysemous nature of the chosen word, leading to distinct interpretations depending on the context.

<sup>2</sup>Beforehand, each video is labeled with the signer identity using the face recognition library Deepface.

To address this issue, when necessary, we categorized the videos according to the word’s specific meaning in the context of each sentence before applying the previous method. To that aim, a Bert language model<sup>3</sup> (Devlin et al., 2019) is used.

**Clustering video clips.** For each query word, a classification is performed on the detected videos in order to discover potential variants. The videos are clustered into classes of similar form. A K-means algorithm is used to that aim and the number of clusters is determined using the Silhouette method (Rousseeuw, 1987). It selects the optimal number of clusters by simultaneously maximizing the distance between clusters and the density of points within each cluster.

Figure 2 shows an example of clustering result for words “Italy” (on the left) and “November” (on the right). For “Italy”, the larger group contains the videos that actually correspond to the sign “Italy”, while the smaller group contains detection errors. For “November” (right), the two groups correspond to two real variants: the two-handed variants on the left, and the one-handed variants on the right.

As the detection errors are automatically grouped during the clustering stage, the use of negative examples to prevent the detection of non-positive frames (section 3.2) may not always be necessary. Nevertheless, we have also employed negative examples for other purposes, as explained below.

<sup>3</sup>Specifically the bert-base-multilingual-cased version from Hugging Face.

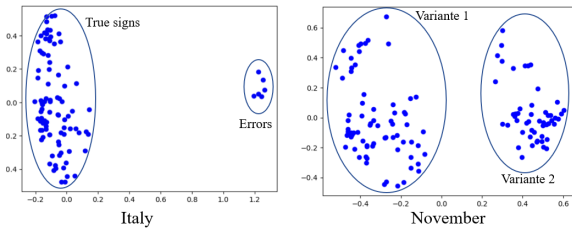


Figure 2: Visualisation of the sequences obtained for “Italy” and “November”. After reducing the size of the data using PCA, the sequences obtained for “Italy” and “November” were projected onto the two main axes.

**Separating frequently associated signs.** For certain words in our list, the model gathers video clips featuring the desired sign alongside another sign. For instance, in Mediapi-RGB, the word “Tokyo” is often associated with “Jeux Olympiques” (Olympic games). To capture only the desired sign, we employed the similarity search of section 3.2 by modifying the set of negative videos. As shown in Figure 3, instead of defining the set of negative videos as those for which the subtitles do not contain the word “Tokyo” (Classic method), the negative samples are defined such as they do not contain Tokyo but contain “Jeux Olympiques” or its equivalent (Custom method).

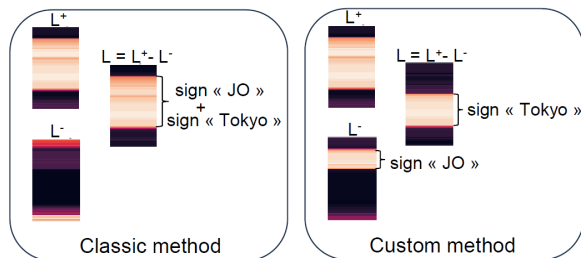


Figure 3: Splitting frequently associated signs using negative examples. In the Custom method (on the right), negative examples are used to locate the sign corresponding to “Jeux Olympiques” in a video for which the subtitle is “Tokyo 2021 Olympic Games are getting closer”. The final vector  $L$  is useful to precisely segment “Tokyo”.

This method is used when precision of the signs segmentation is poor. It proves to be very effective on our data.

After the first stage, a large number of videos clips are extracted from continuous videos. They have various sizes depending on the context and the signer. Each clip is associated with a label based on the word but taking into account any variations in form e.g. juillet\_0, juillet\_1, juillet\_2 (July). As a result, our bilingual dictionary consists of a list of labels to which are associated LSF video clips

containing lexical units.

It is worth noticing that the method is able to discover signs that are not currently available in existing online LSF dictionaries.

In order to assess the quality of the resulting lexicon, a first version of 36 labels is built and audited by LSF experts.

## 4. Step 2: Expert reviewing

The evaluation phase was carried out by two LSF experts. More specifically, the aim was to assess the quality of the segmentation of each clip for each label. To do this, three quality levels are defined:

- 1: when the sign is correctly segmented, that is when it is fully present and there is no frame belonging to the transition parts before or after the sign;
- 2: when it is acceptably segmented, that is some frames belonging to the transitions are present, or a few frames seem to be missing at the beginning or end of the sign;
- 3: otherwise. These are cases where we are able to identify the partial presence of the sign, i.e. it is truncated or accompanied by another sign, possibly not complete. Thus, these occurrences should not be kept for future use.

The choice between categories 1 and 2 is sometimes empirical, typically for signs that include a preparation or retraction phase, which can be blended into the transitions between signs.

Even when the occurrence is perfectly segmented, there may be variations in the shape of the sign, despite the solutions presented in the previous section. We felt it was important to identify the different types of variation so that we could decide whether or not to create separate classes. Three types of variations have been singled out:

- *Lexical*, where there are several signs associated with a given word, for example for certain months such as July.
- *Morphological*, such as the addition of a forward or backward movement with the signs expressing the days of the week, to specify that it is the day of the next or previous week.
- *Internal*, with changes in one of the parameters of the sign (handshape, location, orientation, contact), the number of repetitions or the posture of the dominated arm.

In the first two cases, the form or meaning is different, so separate classes are needed. In the third case, the variations are due to articulatory constraints or individual variants that do not require separate classes.



At the beginning of the process, we had 36 labels with a number of occurrences ranging from 5 to 213. This represented a total of more than 3,000 clips that were manually evaluated by the LSF experts. At the end of the process, we ended up with 53 labels (44 of which had more than 5 LSF examples). Indeed, some of the clusters had to be split because, for example, they contained variants with different meanings (e.g. Wednesday, next Wednesday, previous Wednesday). In addition, because we retained only the occurrences with a 1 or 2 quality level, the total number of video clips has been halved. Therefore, the number of occurrences for each label is lower (from 3 to 202), but the occurrences are more representative. The experiments of section 6 examine how expert enhancement affects classification performance.

## 5. Step 3: Supervised classification

Since we have a French-LSF lexicon (with or without refinement by experts), it becomes possible to design a supervised classification, which will be useful for annotating any continuous LSF video.

### 5.1. Preparing data

For this step, we select from a French-LSF lexicon the labels that have at least 5 LSF examples. Complete videos containing any of these labeled instances are retained. Each frame within these complete videos is assigned to a class label (coded as an integer). However, due to potential missed annotations, some signs may not have been annotated, leading to a partial ground-truth. For example, in a video with the subtitle “Hello, we are Tuesday, April 3rd,” we only captured the sign corresponding to “Tuesday”. The annotation for this video is in the form [00...0066600...00], where 6 is the identifier for “Tuesday”. This annotation is incomplete since the sign corresponding to “April” is not annotated (nor the sign for “Hello”). Therefore, we trained models with data that is partially annotated, making model optimization challenging and quantitative evaluation approximate.

### 5.2. Model Architecture

The system architecture is illustrated in Figure 4:

- The first model extracts video features using a Video Swin Transformer trained on BSL data (the same model as used in Section 3).
- The second model is a lightweight straight-forward MLP classifier. It takes the features as input and produces sequences of integers as output. Each integer in the output sequence identifies the class corresponding to each frame.

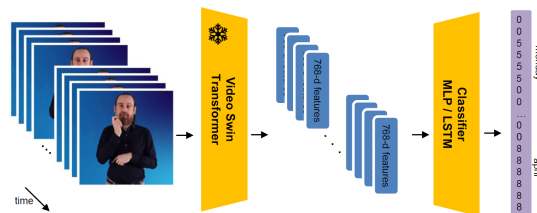


Figure 4: Model Architecture.

### 5.3. Training Setup

The classifier is trained with batches corresponding to non-shuffled features of videos for 15 epochs, using Adam optimization with an initially fixed learning rate of  $1e-4$ . We also used L2 penalty (weight decay =  $1e-5$ ). The Video Swin Transformer was frozen and we initialized the classifier neural network’s weights with Xavier Initialization.

**Loss.** We used the cross-entropy cost function, which is particularly suitable for multi-class classification models. Since the dataset is highly imbalanced (90% of the images are annotated as 0), we applied weights to the cost function. The weights  $w_c$  assigned to each class  $c$  are defined as follows:

$$w_c = 1 - \frac{\text{number of examples for class } c}{\text{total number of examples}}$$

**Metrics.** To assess the quality of the models, we measure accuracy, F1-score, and recall, as follows. First, F1-scores  $F1_c^i$  (or recall  $R_c^i$ ) are computed for each video  $i$  and each class  $c$  present in the ground truth of video  $i$ . For each class  $c$ , these scores are averaged to get  $F1_c$  (or recall  $R_c$ ). The final F1-score (or recall  $R$ ) is finally obtained by averaging the  $F1_c$  (or  $R_c$ ).

As mentioned previously, the ground truth annotation is partial since not all occurrences are identified. However, when annotated, the signs are well segmented and reliable. Therefore, during training, we choose the model with the best recall to minimize the likelihood of missing true positives.

**Sign Classifier.** We tested several architectures of the classifiers, considering both MLPs and LSTMs with one or two layers and hidden layers of 100, 200, or 300 neurons.

In this study, we focus on experiments involving a 2-layer MLP with 200 neurons. We introduced a Normalization layer, used the ReLU activation function after the first layer, and applied a softmax at the output. For the evaluation, a smoothing function is used to eliminate isolated signs.

## 6. Experiments

We carry out several experiments on this model. The first experiments aim to study the impact of the expert analysis and their modification of the dictionary on the supervised annotation. This is made both quantitatively and qualitatively. A second experiment aims to increase the size of the initial vocabulary, in order to evaluate the scalability of our procedure.

### 6.1. Expert versus non expert

This experiment explores the contribution of experts in the data sorting process.

**Data.** In the concluding phase of the initial stage (Section 3), we organized, for each word of our list, a set of automatically clustered videos. Subsequently, these videos underwent a preliminary manual sorting process, involving the removal of clusters corresponding to detection errors and the adding of potential variants. This sorting was carried out by non-experts<sup>4</sup> in a first step, and then by experts (as detailed in Section 4). We consequently obtained a non-expert and an expert dictionaries  $D1$  and  $D2$ , from which we acquired annotated videos (Table 1).

	nb. classes	nb. signs	nb. annot. videos
w/o expert	37	3137	2657
w expert	45	1773	1613

Table 1: Data quantification - w/o and w expertise. In each case, there is an additional class corresponding to a null class.

Note that the scenario involving expertise is more challenging because there are more classes and fewer occurrences per class.

**Quantitative results.** Table 2 presents the results obtained for two classifiers trained with the setup described in Section 5.3, using data sorted with and without expert involvement. In both cases, the data was divided into training, validation and test sets. For consistency, the same videos were selected for the validation and the test set (respectively 227 and 225 videos).

<sup>4</sup>Non-experts: Machine Learning computer scientists who, through working with sign language videos, are presumably capable of comparing sign videos and decide if the signs correspond to the same lexical unit. They do not have the expertise to determine whether a sign will be segmented perfectly, nor to distinguish fine variations of signs.

Data	Recall	F1	Accuracy
w/o expert	0.85 ( $\pm 0.008$ )	0.77 ( $\pm 0.003$ )	0.95 ( $\pm 0.005$ )
w expert	0.85 ( $\pm 0.017$ )	0.78 ( $\pm 0.01$ )	0.95 ( $\pm 0.004$ )

Table 2: Scores on Test set of the classifiers trained on data with and without expertise.

As explained before, the non-expert dictionary  $D1$  contains 36 labels, while the expert one  $D2$  contains 44 labels. The new labels are created by separating variants of form or meaning. In some cases, the differences in the forms can be tricky to perceive, which is why the first automatic step grouped them in a single class.

This is the case for example for the  $D1$  class “mercredi” (Wednesday) that has been split by experts into 3 labels in  $D2$ , which are “mercredi” (Wednesday), “mercredi dernier” (previous Wednesday), “mercredi prochain” (next Wednesday). These three signs with different meanings differ only in the strong hand movement. In SL, time is expressed along the camera axis, with the past to the rear, the present at the level of the signer and the future forwards. What differs on this axis alone is of course more complicated to distinguish in video-type data, which raises a greater challenge to the classifier. However, neglecting this expertise step can lead to major errors which will subsequently have a detrimental effect on task performance.

Thus, our two classifiers are trained on a dictionary  $D1$  with fewer classes and more occurrences, but less precision on both form and meaning, and a dictionary  $D2$  with more classes and fewer occurrences, but more precision on form and meaning. The performance of the two classifiers is very promising and shows that, despite the fact that  $D1$  contains more data, using  $D2$  produces similar scores. There is no difference despite the more challenging conditions of the expertised lexicon and, above all, much greater precision.

**Qualitative analysis.** Figure 5 shows an example of automatic annotation of lexical units on a test video with the subtitle “But the G7 countries - Canada, France, Germany, Italy, Japan, the United Kingdom and the United States - reached an agreement on Saturday”. For this qualitative study, an annotation by a LSF expert has been done on the video using Elan software<sup>5</sup>.

In both cases, all signs are recognized, and are relatively close to the ground truth. Sign segmentation differs slightly between the two classifiers. In this example, the “with expert” classifier is able

<sup>5</sup><https://archive.mpi.nl/tla/elan> - Max Planck Institute for Psycholinguistics (Nijmegen)

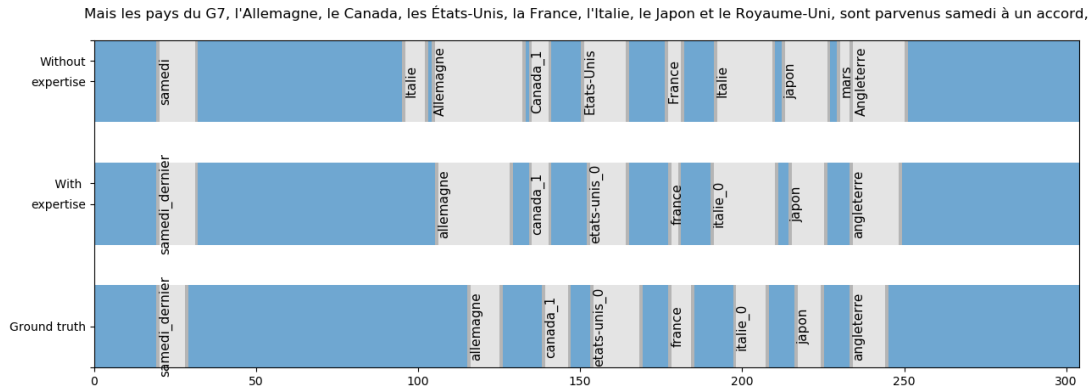


Figure 5: Comparison between the predictions of the non-expert (top), the expert (middle) classifiers and a ground truth (bottom) on a test video.

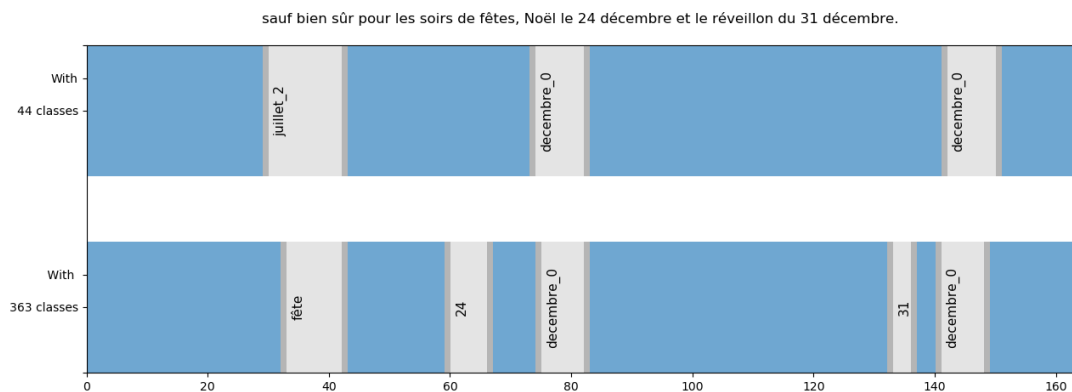


Figure 6: Comparison between the predictions of the 45 and the 364 classes classifier.

to eliminate two insertions present in the version “without expert” classifier (insertions of “Italy” and “March”).

## 6.2. Towards a much larger dictionary

The experiment is extended by increasing the number of the dictionary entries, following these steps:

- Creation of a dictionary comprising 363 labels: 44 sorted by experts (same as in 6.1), to which we added 319 labels sorted by non-experts<sup>6</sup>. In total, 7339 occurrences of signs were collected.
- Annotation of 6047 videos using this dictionary.
- Training of a 364-classes classifier using the training setup described in Section 5.3.

The model achieved an accuracy of 0.93, a recall of 0.65, and a F1-score of 0.63 on the test set.

The figure 6 illustrates the predictions of the expert 45-classifier from Section 6.1 and the predictions of the new 364-classifier on a test video with

<sup>6</sup>The sorting of videos was conducted by non-experts due to time constraints, but we nevertheless believe it would be beneficial for this step to be carried out by experts.

the subtitle: “except, of course, for Christmas Eve on 24 December and New Year’s Eve on 31 December.”

The 364-classifier predicts five positive signs, while the 45-classifier only three. “juillet\_2”, a variant of “juillet” (July), corresponds to the same sign as “fête” (celebration)<sup>7</sup>. This suggests that the 364-dictionary contains two labels for the same form with a different meaning. This is usually not recommended, but the classifier appears to perform well.

## 7. Conclusion and prospects

The paper has presented a system designed for the automatic annotation of lexical units in LSF videos, with an initial vocabulary of 36 labels. This lexicon has been extended to 44 and then 363 labels. Our Pytorch [implementation](#) is publicly available.

The proposed method highlights the transferability of a SL video encoder from one SL (BSL) to another one (LSF).

A non expert dictionary has been compared to

<sup>7</sup>This is due to the celebration of “14 juillet”.

an expert one, in the context of sign recognition in continuous videos. Without expertise, results are very convincing. Yet it hides a problem, which is a lack of precision to distinguish certain signs, notably when they differ according to the motion along the camera axis (e.g. last Wednesday versus next Wednesday). It has shown that, even when using elaborated video encoders such as Video Swin Transformer, not all the subtleties of SL are caught, such as the use of space, which can change the meaning of signs. In our experiments, the expertise has provided a refinement of the classes which is overriding to keep the meaning of the utterances.

Progress is underway, with the next step being to expand the dictionary, coupled with expert review to achieve a vocabulary of 1000 words. The final goal is to annotate the entire Mediapi-rgb corpus as finely as possible, while simultaneously creating a sufficiently large dictionary to train specific video encoding models for LSF.

## 8. Acknowledgements

We would like to thank Media'Pi! for allowing us to use the invaluable bilingual resources they produce for our research. We thanks also Diandra Fabre from Gipsa-Lab for her help with the non-expert data cleaning.

## 9. Bibliographical References

- S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman. 2020. [Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues](#). In *ECCV*, volume 12356, pages 35–53.
- V. Belissen, A. Braffort, and M. Gouiffès. 2020a. [Dicta-Sign-LSF-v2: Remake of a continuous French Sign Language dialogue corpus and a first baseline for automatic sign language processing](#). In *LREC*, pages 6040–6048, Marseille, FR.
- V. Belissen, A. Braffort, and M. Gouiffès. 2020b. [Experimenting the automatic recognition of non-conventionalized units in sign language](#). *Algorithms*, 13(12):310.
- P.J. Buehler, A. Zisserman, and M. Everingham. 2009. [Learning sign language by watching tv \(using weakly aligned subtitles\)](#). *IEEE CVPR*, pages 2961–2968.
- H. Bull. 2023. [Learning sign language from subtitles](#). Ph.D. thesis. Université Paris-Saclay.
- J. Carreira and A. Zisserman. 2017. [Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset](#). pages 4724–4733. *IEEE CVPR*.
- J. Devlin, M-W. Chang, K. Lee, and K. Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Conf. of the NAACL association: Human Language Technologies*, page 4171–4186. ACL.
- T. Jiang, N.C. Camgoz, and R. Bowden. 2021. [Looking for the Signs: Identifying Isolated Sign Instances in Continuous Video Footage](#). In *IEEE FG*, pages 1–8, Jodhpur, India.
- M. Kopf, M. Schulder, and T. Hanke. 2022. [The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources](#). In *LREC Work. on the Repr. and Proc. of Sign Languages*, pages 102–109, Marseille, France.
- D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li. 2020. [Tspnet: Hierarchical feature learning via temporal semantic pyramid for sign language translation](#). In *NeurIPS*, volume 33, pages 12034–12045.
- Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. 2022. [Video swin transformer](#). In *CVPR*, pages 3202–3211, New Orleans, USA. IEEE.
- L. Momeni, H. Bull, K. R. Prajwal, S. Albanie, G. Varol, and A. Zisserman. 2022. [Automatic dense annotation of large-vocabulary sign language videos](#). In *ECCV October 23–27*, page 671–690.
- L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman. 2020. [Watch, read and lookup: learning to spot signs from multiple supervisors](#).
- Y. Ouakrim, D. Beutemps, M. Gouiffès, T. Hueber, F. Berthommier, and A. Braffort. 2023. [A Multistream Model for Continuous Recognition of Lexical Units in French Sign Language](#). In *GRETSI 2023*, Grenoble, France.
- Y. Ouakrim, H. Bull, M. Gouiffès, D. Beutemps, T. Hueber, and A. Braffort. 2024. [Mediapi-rgb: Enabling technological breakthroughs in french sign language \(lsf\) research through an extensive video-text corpus](#). In *20th International Conference on Computer Vision Theory and Applications (VISAPP)*.
- K. R. Prajwal, H. Bull, L. Momeni, S. Albanie, G. Varol, and A. Zisserman. 2022. [Weakly-supervised fingerspelling recognition in british sign language videos](#). In *BMVC*, London, UK.
- K. Renz, N. C. Stache, S. Albanie, and G. Varol. 2021. [Sign language segmentation with temporal convolutional networks](#). In *IEEE ICASSP*, pages 2135–2139, Toronto, Canada. IEEE.



P. J. Rousseeuw. 1987. *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*. *Journal of Computational and Applied Mathematics*, 20:53–65.

H-D. Yang, S. Sclaroff, and S-W. Lee. 2009. *Sign language spotting with a threshold model based on conditional random fields*. *IEEE Trans. on PAMI*, 31(7):1264–1277.

## 10. Language Resource References

Belissen, V. and Braffort, A. and Gouiffès, M. 2020. *Dicta-Sign-LSF corpus*. ISLRN 442-418-132-318-7.

Bull, H. and Ouakrim, Y and Lascar, J. and Braffort, A. and Gouiffès, M. 2024. *Mediapi-RGB corpus*. ISLRN 421-833-561-507-6.

# Capturing Motion: Using Radar to Build Better Sign Language Corpora

Evie Malaia , Joshua Borneman , Sevgi Gurbuz 

University of Alabama, Purdue University, University of Alabama  
Tuscaloosa, AL; West Lafayette, IN; Tuscaloosa, AL  
eamalaia@ua.edu, jdbornem@purdue.edu, szgurbuz@eng.ua.edu

## Abstract

Sign language conveys information using dynamic visual signal. Proficient signers rely on the skill in processing and predictive motion information during sign language comprehension. Much current work in sign language corpora development relies on video data. However, from the perspective of information transfer in communication, video recordings are limited in capturing spatial and temporal frequencies of sign language signal in sufficient resolution. In contrast, radar can capture 3D motion data at high temporal and spatial resolution, preserving depth articulations lost in 2D video. Radar's recording parameters can also be adapted in real time to optimize temporal resolution for rapid signing motions. Thus, radar recordings provide higher-fidelity corpora for analyzing linguistic features of sign languages and creating smart environments that respond to signed input. Crucially, radar recordings uphold user privacy, only capturing kinematic parameters of communicative signal, as opposed to signer identity. Radar resolution in capturing dynamic data from sign language production, and privacy advantages it provides to users, make it uniquely suited for advancing sign language research through corpora development.

**Keywords:** sign language, production, radar

## 1. Signed Communication

Sign languages convey linguistic information dynamically through articulator motion. Although linguistic analyses of signs only identifies motion as a component of sign phonology, on par with hand-shape, hand orientation, and place of articulation, research in visual perception and sign comprehension has long been clear on relevance of dynamic motion, as opposed to static components of articulation, to proficient signers (Malaia et al., 2023). Lifelong exposure to visual complexity inherent in sign motion affects both perceptual and cognitive processing in sign language users compared to non-signers, and enhances signers' perceptual tuning to the information density in motion signals, allowing them to parse continuous signal, identifying discrete signs and their grammatical modifications (Klima et al., 1999; Bavelier et al., 2006). Linguistic distinctions in meaning and grammar are reflected in the movement dynamics of the signed signal. These distinctions can be captured in a manner parallel to acoustic and phonetic analysis of spoken signals (Borneman et al., 2018).

Fully visible articulator motion in sign language carries all communicated information. At the same time, sign language motion carries more information defined as visual signal entropy than everyday human motion (cf. Fig. 1). The parameters that are critical to capturing information-dense features of the continuous signal are the temporal resolution and the amount of change present in the signal within the given time window. When signs are produced fluently in sentences, there are almost

always transitional movements between them, for example, when one sign ends with the hand(s) located in one place and the next sign starts with them located somewhere else, there must be a movement of the hands to that next location before the next sign can start its lexical movement. These transitions are clearly differentiable to signers, and ignored when they are asked to count/tap to syllables (Klima et al., 1999).

The variability of motion between sign-syllables and transitions forms the basis of the quantitative distinction between non-informative, biological motion, and the sign language signal (Malaia et al., 2018). Mathematically quantified amount of information (i.e. variability) in the motion of the articulators in sign language forms the basis of sign syllables (Malaia and Wilbur, 2020). Experimental approaches, including video analysis using optical flow and motion capture data analysis, indicate that information transfer in sign language critically relies on the entropy of the articulator signal, making it critical to capture dynamic changes in it with sufficient spatial and temporal resolution.

## 2. Information Transfer in Sign Language Signal

When evaluating and comparing modalities for capturing sign language motion, and for analyzing languages in general, a key factor is the fidelity and dimensionality with which each modality can capture the information content of the original dynamic signal over time (Malaia et al., 2022). It is first useful

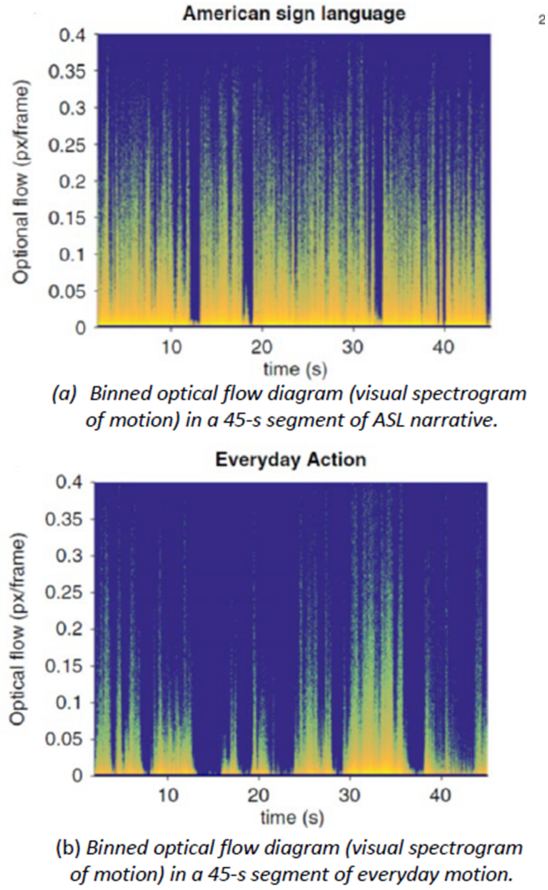


Figure 1: ASL and action: comparative variability optical flow spectrograms (a - American Sign Language; b - everyday motion).

to explain the common framework on which different language types, and capture methods, may be compared.

Starting with a simple example, a spoken language is a 1-dimensional time series signal, carrying information in amplitude as a function of temporal frequency  $[f_t]$ , written here as  $[S_0(f_t)]$ . Recording of this spoken signal, usually limited in amplitude and frequency by electronics/sampling method, may then be treated as a series of transfer functions. For instance we may have recording/electronics/sampling effects,  $[T_r(f_t)]$ , and effects on the data due to preprocessing  $[T_p(f_t)]$ . Importantly,  $T(f_t) < 1$ , i.e. no recording or capture method is perfect. This means that the final recorded language signal is not a pure recorded sample of the original spoken language, but is rather a modified signal  $[S_1(f_t)]$ , where  $S_1(f_t) = S_0(f_t) \cdot T_r(f_t) \cdot T_p(f_t)$ . Therefore, final analysis of the spoken language is always done on a reduced fidelity recording. Knowing this, a spoken language recording method should be selected which preserves the overall information density within the

temporal component,  $[f_t]$ . This would require high temporal sampling and analysis frequencies, and most acoustic recordings may contain a minimum of 20k samples per second. This characterization may seem trivial for a 1-dimensional spoken language, but the framework becomes useful when dealing with a multi-dimensional signal, such as sign language. Compared to linear sound recordings, capture of sign language presents a significant difficulty. Sign language conveys information over spatial frequencies in 3 space dimensions  $(f_x, f_y, f_z)$ , as well as in temporal frequency  $(f_t)$ , and therefore any analysis of sign languages will depend on the accuracy and dimensionality with which the original signal can be recorded, as well as potential dimensionality reduction and fidelity loss during further analysis. Each recording and processing step acts as a filtering function, potentially reducing the fidelity of the data. Therefore, it is important to select measurement and analysis methods which preserve, or at least are intentional about, how dimensionality and fidelity are addressed. Sign language, as a 3-dimensional spatial signal also varying in time,  $S_0(f_x, f_y, f_z, f_t)$ , is filtered in both spatial frequencies,  $f_x, f_y, f_z$ , and temporal frequencies,  $f_t$ , depending on how it is recorded  $[T_r]$  and how it is preprocessed  $[T_p]$ .  $S_1(f_x, f_y, f_z, f_t) = S_0(f_x, f_y, f_z, f_t) \cdot T_r(f_x, f_y, f_z, f_t) \cdot T_p(f_x, f_y, f_z, f_t)$  Although sign languages use relatively lower temporal frequencies as compared to spoken language, sign language also transfers information in additional spatial dimensions. These spatial dimensions must also be recorded in order to preserve the overall information density. This description may now be used to describe various methods of language capture in a common framework. For example, video capture recordings of sign language are, in essence, a 2D spatial frequency filter, which removes depth information  $[T_r(f_z) = 0]$ , and in which the  $x, y$  spatial plane is downsampled to  $s, t$  by the camera distance and resolution  $[(f_u, f_v) \approx (f_x, f_y)]$ , and filtered such that  $[T_r(f_u, f_v) < 1]$ . The camera resolution and position should ideally be placed such that all hand/arm articulators in the signing space are resolved, that is, that the articulator frequencies are in the camera band-pass. Further,  $f_t$  is subsampled by the frame rate of the video recording  $[f_T \approx f_t]$ , resulting in  $T_r(f_T) < 1$ . Therefore, our pure real-world sign language information signal  $[S_0(f_x, f_y, f_z, f_t)]$  is now recorded by video and subsampled to only two spatial dimensions and time  $[S_{1,video}(f_s, f_t, f_T)]$ .

In contrast, radar is capable of capturing 3D motion data over time, with adaptive temporal resolution based on user-configurable recording parameters. Radar signal processing algorithms may extract range-Doppler (RD) maps (2D images of range versus Doppler frequency) or micro-Doppler

signature (Doppler frequency versus time). Therefore, radar records motion along the depth axis  $z$ , subsampled to  $w$  resolution [ $f_w \approx f_z$ ], such that  $[T_r(f_w) < 1]$  through the line of sight distance. The remaining spatial dimensions  $f_x, f_y$  are convoluted into a radial velocity and angle of arrival such that  $[(f_r, f_a) \propto (f_x, f_y)]$  and therefore  $[T_r(f_r, f_a) < 1]$ . Temporal resolution is adjustable based on the pulse repetition frequency (PRF), and can be set to match sign language motion bandwidths, and is generally faster than video frame rates,  $[T_r(f_T) < 1]$ . For Frequency Modulated Continuous Wave (FMCW) radar, the PRF also determines the maximum measurable radial velocity ( $v_{max} = PRF \times \lambda/2$ ) and the velocity resolution  $\Delta v = PRF/N$ , where  $N$  is the total number of pulses transmitted. With higher transmit frequencies, the Doppler frequency shift incurred due to even slower motions is greater and hence more easily measurable; however, this also reduces the maximum velocity limit. Thus, selecting a high PRF is advantageous both from the perspective of ensuring unaliased velocity measurements and high temporal sampling of motion during signing. In prior work comparing the resulting radar micro-Doppler signatures of lower bandwidth ( $\beta$ ), lower PRF signal with low duty cycle ( $d$ ) ( $\beta = 750$  MHz, PRF = 3.2 kHz,  $d = 51.2\%$ ) versus one of high bandwidth, PRF and duty cycle ( $\beta = 4$  GHz, PRF = 6.4 kHz,  $d = 96\%$ ), we found that the latter enabled crisp and pristine micro-Doppler signatures of sign language (Gurbuz et al., 2022a). Spatial depth resolution depends on the transmitted waveform's bandwidth as  $\Delta r = c\beta/2$ , where  $c$  is the speed of light. Ideally, an FMCW radar with high bandwidth and high PRF is best suited for sign language measurements, as this enables both high spatial and temporal resolution measurements. Automotive radars are well-suited for this aim as they typically have bandwidths of 4 GHz and are designed with PRFs so high as to measure vehicular speed. As the commercial applications of low-cost, low-power radar sensors are ever expanding, it is now possible to find sensors having bandwidths of 5 or even 7 GHz. The main disadvantage of operating the sensor at such high bandwidth and PRF is the high volume of data that results from high spatiotemporal sampling. However, such considerations can be mitigated by interactively adapting the transmission parameters of the waveform so that a low spatiotemporal resolution waveform is transmitted when no human presence is detected, or if a person is simply engaging in daily activities, while a high spatiotemporal resolution waveform is transmitted once a device is triggered and sign language recognition is needed (Kurtoglu, 2024).

Therefore, our pure real-world sign language information signal  $[S_0(f_x, f_y, f_z, f_t)]$  is sub-sampled

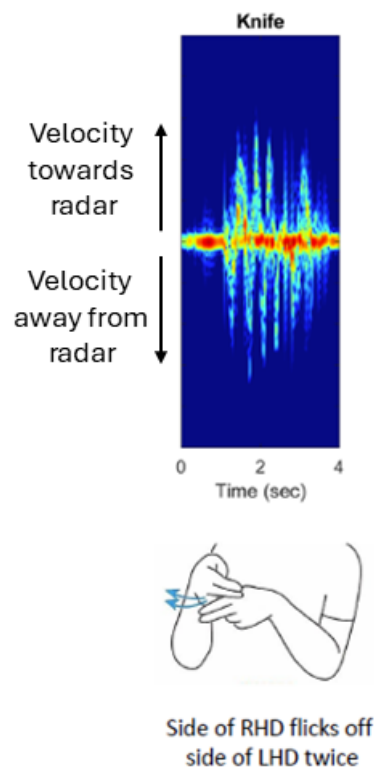


Figure 2: Sample radar micro-Doppler signature for the sign KNIFE.

with radar to two convoluted spatial dimensions, one pure spatial dimension, and time  $[S_{1,radar}(f_v, f_a, f_z, f_t)]$ . This dimensional analysis is useful to evaluate not just hardware capture, but signal processing (Malaia et al., 2022). However it is seen here that compared to 2D video, radar provides crucial depth information about sign articulations in 3D space. Radar's recording parameters can also be selected to maximize temporal resolution appropriate for capturing the rapid motions of signing - a PRF of 6.4 kHz, as utilized in our earlier example, offers much higher temporal sampling than that of a high-speed webcam, which can have a frame rate of about 200-300 frames per second (fps). Thus, the micro-Doppler signature offers a novel representation of sign language corpora that can capture sign language kinematics in a unique fashion, while also doing so in an ambient fashion without recording private imagery. Consider, for example, the micro-Doppler signature for the sign KNIFE, shown in Figure 2. Not only can the maximum and minimum velocity in both directions be measured, but also the timing of the repetitive motion and the number of times the fingers moved back and forth. Notice also that from the radar image we cannot infer any information about the location or environment the recording was made or even who was signing.



In addition to manual articulations, sign languages also involve facial expressions, mouth shapes, contact between the fingers and body, as well as eye movements, which hold linguistic significance. These are areas of ongoing, active research in radar technology, which may one day make radar-based sign language studies beyond manual articulations possible. For example, lip reading under face masks using radar has been proposed (Hameed et al., 2022) to enable speech recognition when camera-based techniques are not possible due to the obstruction by the mask. Emotion recognition (Dang et al., 2022) has also become a topic of interest, as such facial movements during expressions is coupled with vital signs recorded by the radar. Moreover, through the use of a high number antenna array elements in both the azimuth and elevation, newly developed high-resolution imaging radars (Bräunig et al., 2023) have been developed that can provide a distinct image of hand shape, which can thus enable recognition of fingerspelling. The principle downside of this current technology, however, is that such imaging radars are not able to dynamically acquire images and require the hand to be stationary. However, as automotive radars are commercialized with an increasing number of array elements, so is the azimuth and elevation angular resolution increasing so that potentially new AI/ML algorithms operating directly on the raw radar data can be developed to enable such functions that require high spatial resolution and localization (such as detection of finger-body contact).

For these reasons, radar provides a uniquely informative way for capturing sign language corpora, which we have only yet begun to explore. Radar's higher-fidelity 3D motion data over time offers potential for more detailed analysis of linguistic and kinematic features of sign languages. This advantage highlights radar's promise for advancing sign language research through improved corpora.

### 3. Radar-based Sign Corpora and Machine Learning

Unlike video, radar measurements are not inherently an image, but are actually a time-stream of complex I/Q data from which line-of-sight distance, radial velocity, and angle of arrival may be computed. The radar measurements may be visualized via a variety of 2D and 3D data tensors. The most widely used representation is the *micro-Doppler signature* (Chen, 2019), which is computed using the short-time Fourier transform and reveals the micro-Doppler frequencies - or radial velocity - due to rotational motion centered about the Doppler shift due to translational motion. Thus, the micro-Doppler signature is a rich source of kinematic information relating to signing. In our prior work

(Rahman et al., 2022), we have shown that using micro-Doppler signatures alone, snapshots of over 100 word-level signs can be classified at over 77% top-1 and 92% top-5 accuracy. Moreover, we found that RF micro-Doppler frequencies also captured significant linguistic properties of the signer, such as co-articulation (Gurbuz et al., 2020), whether the signer was fluent in ASL versus being a hearing imitation signer (Gurbuz et al., 2021, 2022b), and whether or not the signer was being directed to articulate a sign versus doing a natural articulation as part of freely playing a game (Kurtoglu et al., 2024).

Thus, the linguistic characteristics of a signer have a significant impact on model training: deep neural networks (DNNs) for recognition of natural, fluent signing cannot be effectively trained using imitation signing or signing acquired via controlled experiments. Integration of kinematic constraints into the DNN architecture itself is also greatly beneficial. For example, the envelopes of the micro-Doppler signature measure the peak radial velocity incurred during signing and can be provided as a dual input to the discriminator of a Generative Adversarial Network (GAN) and used to compute a physics-based loss function, which combine enable to GAN to synthesize kinematically more accurate data for model training (Rahman et al., 2022, 2023). The utilization of multi-task learning where loss functions for each task are defined based on kinematic properties is also beneficial for recognition performance (Kurtoğlu et al., 2022). For example, we showed that a trigger sign (or wake word) could be more effectively recognized if the DNN optimized for five distinct tasks: 1) one versus two handedness, 2) major location of hands, 3) movement type, 4) daily activity versus ASL, and 5) number of strokes comprising the sign. Linguistic metrics, such as fractal complexity, were also found to be indicative of whether a person was signing versus doing an everyday activity at home (Gurbuz et al., 2020).

In addition to the micro-Doppler signature, 3D RF data tensors can be used to provide an enriched input to DNNs and achieve greater accuracy when trying to recognize sign language in a real-world environment, such as would occur if a user were to use sign language to interact with an electronic personal assistant, such as Alexa or Siri. Using radar signal processing, the raw radar data stream can be converted into a time series images of range-velocity and range-angle. Joint utilization of multiple radar data representations has been used to design a Joint Domain Multi-Input Multi-Task Learning (JD-MIMTL) network (Kurtoğlu et al., 2022) that can automatically segment and extract signing sequences from continuous recordings of daily life, detect whether a trigger sign has been articulated, and recognize subsequent signs as device com-

mands. In fact, estimation of the angle at which a person is located relative to the location of the radar can be used to generate an angular projection of the RF data tensor for the left and right hands (Kurtoğlu et al., 2023). A multi-view DNN was designed to leverage the separate projections of the left and right hand for increased sign recognition performance.

A major challenge to deep learning based ASL recognition with both video and radar remains the limited availability of data that truly captures the nuanced variations of natural signing. To overcome this challenge, an interactive game (Kurtoglu et al., 2024), ChessSIGN, was developed that acquires both video and radar data as a user articulates ASL to move the pieces of the chess game. When the user clicks on a piece, different ASL words corresponding to valid chess moves appear on the screen. The piece moves its position based on recognition of the user's articulation of the sign. We have shown that for both video and radar data, machine learning models trained under data collected via controlled experiments is not effective in recognizing signing in such an unconstrained, natural setting. However, as the system acquires more and more natural signing data during the course of the game, recognition accuracy increases. Moreover, the signs recorded are natural language recordings, which more accurately reflect 1) variations in ASL due to person-specific traits, regional dialects, and fluency; and 2) natural effects such as coarticulation, which occur due to the variation in the position with which a sign can begin or end, as typical of daily life. ChessSIGN thus provides an entertaining way to minimize the burden on the Deaf community to acquire ASL data, while also continually building improved models. Also, because the system captures simultaneous recordings of video with radar, this unique dataset can enable the exploration of new ASL recognition algorithms that jointly exploit the strengths of radar and video together.

Ultimately, our work has shown that RF sensing can capture the kinematics of the rapid progression of dynamic sign sequences that is characteristic of ASL usage. We not only bring to bear, for the first time, a linguistic perspective to RF-based motion recognition, but also a physics-based machine learning approach achieved through the integration of kinematics with deep learning. These advances have enabled the development of RF-sensing based ASL-sensitive human computer interaction (HCI) and as a tool for linguistic analysis of ASL.

#### 4. Ethical Consideration for Sign Language Corpora

Collecting sign language data with radar sensors also offers important privacy advantages over video recording. Video cameras capture detailed visual information about a person's appearance, clothing, surroundings, and any visible actions. This raises significant personal privacy concerns, especially when recording in homes or private spaces. In contrast, radar does not actually record images or videos. Radar sensors operate by transmitting electromagnetic waves and analyzing the reflected signals. The sensors only measure the time-varying position and velocity of body parts as they move through space. No identifying visual features are recorded. The raw radar data itself reveals nothing about a person's identity, gender, attire, or environment. While video provides full visual details, this level of information is unnecessary for analyzing sign language gestures. The intricate motions of signing are characterized by the changing spatial relationships and dynamics of the hands, arms, and face over time. Radar captures exactly these articulatory parameters relevant to sign language, without any personal identifying visuals. Participants are also more comfortable being recorded by radar since their privacy is protected. No video footage exists that could be leaked or exploited. Radar enables collecting natural, unrestrained sign language data even in private real-world environments. Radar recordings capture information-bearing motion from sign language signal with fidelity sufficient for both linguistic or ML-based analysis, while upholding signers' privacy. The ability to gather realistic sign language data in a completely private manner makes radar systems uniquely suited for building sign language corpora and recognition datasets in an ethical, non-invasive way.

#### 5. Bibliographical References

- Daphne Bavelier, Matthew WG Dye, and Peter C Hauser. 2006. Do deaf individuals see better? *Trends in cognitive sciences*, 10(11):512–518.
- Joshua D Borneman, Evie A Malaia, and Ronnie B Wilbur. 2018. Motion characterization using optical flow and fractal complexity. *Journal of Electronic Imaging*, 27(5):051229.
- Johanna Bräunig, Vanessa Wirth, Christoph Kamel, Christian Schüßler, Ingrid Ullmann, Marc Stamminger, and Martin Vossiek. 2023. An ultra-efficient approach for high-resolution mimo radar imaging of human hand poses. *IEEE Transactions on Radar Systems*, 1:468–480.

- Victor C Chen. 2019. *The micro-Doppler effect in radar*. Artech house.
- Xiaochao Dang, Zetong Chen, and Zhanjun Hao. 2022. Emotion recognition method using millimetre wave radar based on deep learning. *IET Radar, Sonar & Navigation*, 16(11):1796–1808.
- Sevgi Z. Gurbuz, Ali C. Gurbuz, Evie A. Malaia, Darrin J. Griffin, Chris Crawford, M. Mahbubur Rahman, Ridvan Aksu, Emre Kurtoglu, Robiulhossain Mdrafai, Ajaymehul Anbuselvam, Trevor Macks, and Engin Ozcelik. 2020. A linguistic perspective on radar micro-doppler analysis of american sign language. In *2020 IEEE International Radar Conference (RADAR)*, pages 232–237.
- Sevgi Z Gurbuz, Ali Cafer Gurbuz, Evie A Malaia, Darrin J Griffin, Chris S Crawford, Mohammad Mahbubur Rahman, Emre Kurtoglu, Ridvan Aksu, Trevor Macks, and Robiulhossain Mdrafai. 2021. American sign language recognition using rf sensing. *IEEE Sensors Journal*, 21(3):3763–3775.
- Sevgi Z. Gurbuz, M. Mahbubur Rahman, Emre Kurtoglu, Evie Malaia, Ali Cafer Gurbuz, Darrin J. Griffin, and Chris Crawford. 2022a. [Multi-frequency rf sensor fusion for word-level fluent asl recognition](#). *IEEE Sensors Journal*, 22(12):11373–11381.
- Sevgi Z Gurbuz, M Mahbubur Rahman, Emre Kurtoglu, Evie A Malaia, Ali Cafer Gurbuz, Darrin J Griffin, and Chris Crawford. 2022b. Multi-frequency rf sensor fusion for word-level fluent asl recognition. *IEEE Sensors Journal*, 22(12):11373–11381.
- Hira Hameed, Muhammad Usman, Ahsen Tahir, Amir Hussain, Hasan Abbas, Tie Jun Cui, Muhammad Ali Imran, and Qammer H. Abbasi. 2022. Pushing the limits of remote RF sensing by reading lips under the face mask. *Nature Communications*, 13(1):1–9.
- Edward S Klima, Ovid JL Tzeng, YYA Fok, Ursula Bellugi, David Corina, and Jeffrey G Bettger. 1999. From sign to script: Effects of linguistic experience on perceptual categorization. *Journal of Chinese Linguistics Monograph Series*, pages 96–129.
- E. Kurtoglu. 2024. *Fully-Adaptive RF Sensing for Non-Intrusive ASL Recognition via Interactive Smart Environments*. Ph.D. thesis, Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL.
- Emre Kurtoglu, Kenneth DeHaan, Caroline Kobek Pezzarossi, Darrin J. Griffin, Chris Crawford, and Sevgi Z Gurbuz. 2024. Interactive learning of natural sign language with radar. *IET Radar Sonar and Navigation*.
- Emre Kurtoğlu, Sabyasachi Biswas, Ali C. Gurbuz, and Sevgi Zubeyde Gurbuz. 2023. Boosting multi-target recognition performance with multi-input multi-output radar-based angular subspace projection and multi-view deep neural network. *IET Radar, Sonar & Navigation*, 17(7):1115–1128.
- Emre Kurtoğlu, Ali C. Gurbuz, Evie A. Malaia, Darrin Griffin, Chris Crawford, and Sevgi Z Gurbuz. 2022. Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces. *IEEE Transactions on Human-Machine Systems*, 52(4):699–712.
- Evie A Malaia, Joshua D Borneman, Emre Kurtoglu, Sevgi Z Gurbuz, Darrin Griffin, Chris Crawford, and Ali C Gurbuz. 2022. Complexity in sign languages. *Linguistics Vanguard*, 9(s1):121–131.
- Evie A Malaia, Joshua D Borneman, and Ronnie B Wilbur. 2018. Information transfer capacity of articulators in american sign language. *Language and Speech*, 61(1):97–112.
- Evie A Malaia, Sean C Borneman, Joshua D Borneman, Julia Krebs, and Ronnie B Wilbur. 2023. Prediction underlying comprehension of human motion: an analysis of deaf signer and non-signer eeg in response to visual stimuli. *Frontiers in Neuroscience*, 17:1218510.
- Evie A Malaia and Ronnie B Wilbur. 2020. Syllable as a unit of information transfer in linguistic communication: The entropy syllable parsing model. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1):e1518.
- Mohammad Mahbubur Rahman, Evie A. Malaia, Ali Cafer Gurbuz, Darrin J. Griffin, Chris Crawford, and Sevgi Zubeyde Gurbuz. 2022. Effect of kinematics and fluency in adversarial synthetic data generation for asl recognition with rf sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4):2732–2745.
- Mohammed Mahbubur Rahman, Sevgi Z Gurbuz, and Moeness G Amin. 2023. Physics-aware generative adversarial networks for radar-based human activity recognition. *IEEE Transactions on Aerospace and Electronic Systems*, 59(3):2994–3008.

# Exploring Latent Sign Language Representations with Isolated Signs, Sentences and In-the-Wild Data

Fredrik Malmberg<sup>1</sup> , Anna Klezovich<sup>1</sup> , Johanna Mesch<sup>2</sup> , Jonas Beskow<sup>1</sup> 

<sup>1</sup>Division of Speech, Music and Hearing, KTH, Sweden

<sup>2</sup>Department of Linguistics, Stockholm University, Sweden

fmalmb@kth.se, annkle@kth.se,

johanna.mesch@ling.su.se, beskow@kth.se

## Abstract

Unsupervised representation learning offers a promising way of utilising large unannotated sign language resources found on the Internet. In this paper, a representation learning model, VQ-VAE, is trained to learn a codebook of motion primitives from sign language data. For training, we use isolated signs and sentences from a sign language dictionary. Three models are trained: one on isolated signs, one on sentences, and one mixed model. We test these models by comparing how well they are able to reconstruct held-out data from the dictionary, as well as an in-the-wild dataset based on sign language videos from YouTube. These data are characterized by less formal and more expressive signing than the dictionary items. Results show that the isolated sign model yields considerably higher reconstruction loss for the YouTube dataset, while the sentence model performs the best on this data. Further, an analysis of codebook usage reveals that the set of codes used by isolated signs and sentences differ significantly. In order to further understand the different characters of the datasets, we carry out an analysis of the velocity profiles, which reveals that signing data in-the-wild has a much higher average velocity than dictionary signs and sentences. We believe these differences also explain the large differences in reconstruction loss observed.

**Keywords:** sign language data, VQ-VAE, Representation Learning, Pose Codebook

## 1. Introduction

Sign languages play a critical role in the communication of deaf communities worldwide, with over 300 different sign languages in use. Despite their significance, sign languages are generally under-resourced compared to spoken languages, with small corpora and limited lexicon due to the need for a manual gloss annotation of sign language videos. While processing of written and spoken languages has advanced rapidly in recent years, with technology performing on par with humans, the same trend has not yet been observed in sign language processing.

Recent progress in speech and text processing has been possible thanks to self-supervised representation learning methods that can be carried out on vast corpora without the need for manual annotation. It has been shown for a speech generation task that learning a powerful data representations significantly improves speech generation (Baevski et al. (2020), van den Oord et al. (2016)). Importantly, this has also made it possible to train models not only on data specifically recorded for the purpose of language technology, but also on in-the-wild data from various Internet sources such as YouTube, which is very beneficial for the low-resourced domain of sign languages.

In this paper, we are investigating how a Vector Quantized Variational Autoencoder (VQ-VAE) representation learning model can learn a code-

book of motion primitives from pose-tracked video data. We train this model both on dictionary signs and short sentences, and we investigate how the model's performance generalizes to sign language data from YouTube. Examples of sequences reconstructed from the model can be seen on our project page<sup>1</sup>.

In the future perspective this model can be used for producing sign language data representations that can be used as a stepping stone for the sign language generation task.

## 2. Related Work

Unsupervised representation learning has been found effective in various data domains, for example using masked language for natural language understanding tasks (Devlin et al., 2018) or audio pre-training for speech recognition (Baevski et al., 2020). For generation tasks in the motion domain, different kinds of probabilistic representation learning schemes, such as VQ-VAEs have been successful. For instance, in the co-speech gestures domain, Yazdian et al. (2022) paper focuses on learning representations for motion primitives with the help of denoising autoencoder (DAE) model that encodes poses into simpler representations, and then these representations are fed as sequences into the second model – VQ-DVAE, that learns mo-

<sup>1</sup>[www.speech.kth.se/research/vq-sign](http://www.speech.kth.se/research/vq-sign)



tion primitives. In the dance generation domain, [Siyao et al. \(2022\)](#) paper uses VQ-VAE as a step for dance generation. The VQ-VAE learns choreographic motion primitives and then an actor-critic GPT model generates dances out of the motions coherently with the music. For more general motion, [Jiang et al. \(2023\)](#) trains a VQ-VAE to create a motion vocabulary that is then used together with a GPT model for several tasks such as Text-to-Motion, Motion Prediction and Motion-to-Text.

Recently, similar representation learning approaches have been applied also to sign language data for sign language understanding task, e.g. a SignBert paper by [Zhou et al. \(2021\)](#) and newer SignBERT+ by [Hu et al. \(2023\)](#). More specifically, a VQ-VAE model has been applied to sign language in [Xie et al. \(2022\)](#), where the main focus is on sign pose sequence generation using a diffusion model. However, in [Xie et al. \(2022\)](#) the authors encode poses frame by frame in the latent space, and as a result they get encoded key points per frame instead of motion primitives capturing a sequence of frames. In our work we use VQ-VAE as a way to learn a codebook of motion primitives for sign language data.

### 3. Data

#### 3.1. Swedish Sign Language Dictionary

This study uses the Swedish Sign Language (STS) Dictionary [Svenskt teckenspråkslexikon \(2024\)](#), which contains 21 000 entries and 6700 sentence examples. Each dictionary entry includes a video of the sign, phonological information, variants, and example sentences. The Swedish Sign Language Dictionary is also linked to the Swedish Sign Language Corpus through ID-glosses ([Mesch et al., 2012](#); [Mesch and Wallin, 2015](#)). It highlights how this focuses on lexical issues, particularly sign lemmatization, and aims to offer a more comprehensive lexical description and understanding of language use in natural conversation settings. The total duration of the dictionary data is 664 minutes, or 1 731 976 frames.

#### 3.2. YouTube Data

For the purposes of testing our representation learning model on the in-the-wild data, we collected data from the YouTube channel "UR Teckenspråk"<sup>2</sup>. Our YouTube dataset contains 17 videos from the Djupdyk playlist with a total duration of 105.6 minutes and a total number of frames 158 406, which is comparable to the size of our test dataset (99 minutes and 229 802 frames respectively).

<sup>2</sup>[www.youtube.com/@URTeckensprak](http://www.youtube.com/@URTeckensprak)

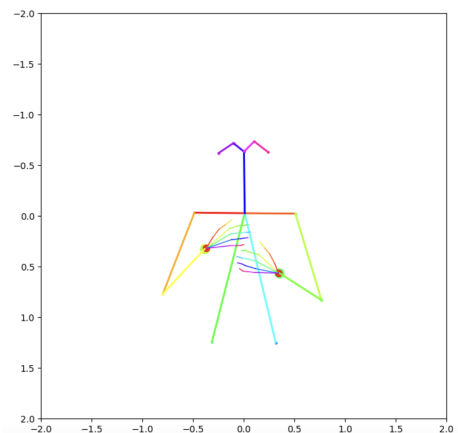


Figure 1: Example of extracted and normalized keypoints. Hands are relocated to wrist positions and lines are drawn between keypoints for illustration purposes.

#### 3.3. Pose Tracking and Pre-Processing

DW Pose ([Yang et al., 2023](#)) was used to extract 2D pose keypoints frame by frame in the videos. The decision to use DW Pose over the commonly used MediaPipe ([Zhang et al., 2020](#)) was based on its subjectively perceived robustness for the specific use case. For the sake of simplicity, only keypoints relating to the overall upper body pose, arms and hands were used resulting in 56 2D keypoints per frame (see Figure 1 for an example). In the future perspective, we want to add facial features since most non-manuals are carried out through the facial features.

In order to preprocess the raw pose data, we select the center of the first frame (the keypoint that connects body with the neck) in the sequence in order to shift the bodypose with respect to it, and then we scale the pose by a scaling factor based on the distance between the left and right shoulder. The keypoints related to the hands are shifted so that the wrist is located in the center for each frame in order to capture finger movements and hand shapes regardless of their global position.

#### 3.4. Velocity Profile Examination

Our VQ-VAE model architecture requires choosing the sequence length to encode. Since we wanted to find motion primitives for sign language data, we investigated velocity profiles of the STS dictionary dataset to estimate the appropriate sequence length to encode.

In order to find velocity, we calculated centroids for each hand coordinates and then computed the distance between the centroids of neighboring video frames separately for each hand. Velocity was calculated as an Euclidean norm of the

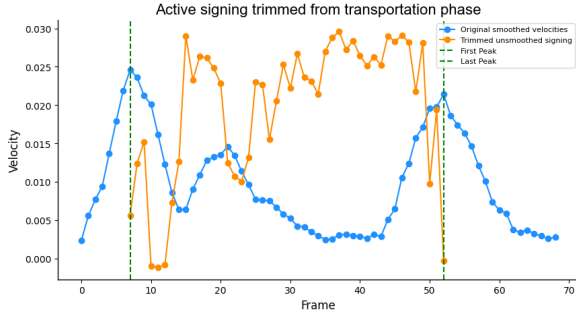


Figure 2: Velocity profile for the STS sign 'kloster'. First peak and last peak signify the transportation phase. Velocity calculated as a distance between hand coordinates centroids for neighboring frames.

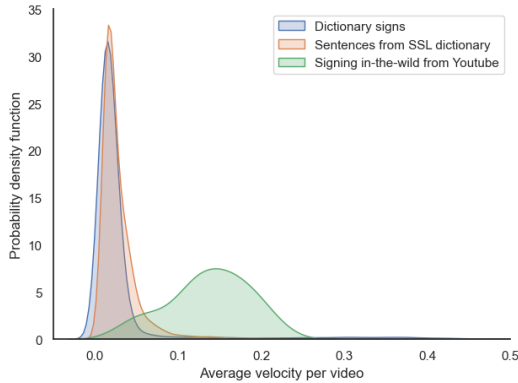


Figure 3: Comparison of average velocity distributions for three types of data.

displacement vector between the centroids and averaged between hands, because we have both left handed and right handed signers in the dataset.

By studying the the number of velocity peaks between preparation for signing and retraction movements, we found that an average number of frames that correspond to one motion in both dictionary signs and sentences is around 30 frames. Similarly to Börstell (2023) we used the moving average to smooth the signal and extract the first and last peaks that correspond to transportation movements (preparation and retraction). For the analysis, we trimmed a signal from transportation movements, and then extracted the peaks from the inverted active signing signal based on a heuristic where the peak is significant if it is higher than one standard deviation from the mean (see example of a velocity profile for sign 'kloster' in Figure 2). As a result, we estimated that the average number of frames for motions in a sentence is 31.7 frames and 26.3 in dictionary signs.

This information was used then in the model design stage, where we assigned sequence length in the VQ-VAE to 30 frames for both signs and phrases based models.

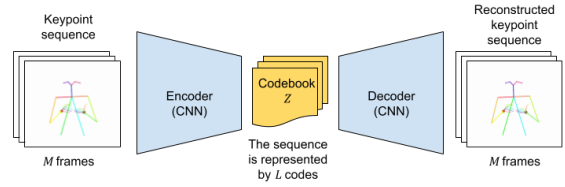


Figure 4: The VQ-VAE consists of an encoder that takes sequences of poses as inputs, a codebook that captures the motion codes and a decoder that outputs reconstructed sequences.

Additionally, we compared the distributions of average velocities for videos in three datasets that we are using (see Figure 3). As a result, we discovered that signing in the wild is much faster than both dictionary sentences and signs, as expected. While the velocity of signing in sentences from a dictionary is only a little bit higher than the velocity of dictionary signs. We expected the velocity of sentences to be closer to the signing-in-the-wild.

## 4. Model

### 4.1. VQ-VAE Architecture

Inspired by the architecture in Jiang et al. (2023), that focuses on tokenizing body motion, we train a 2D sign motion tokenizer using a VQ-VAE (van den Oord et al., 2018). It consists of an encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$ , with a discrete latent representation transforming motion into a structured codebook.  $\mathcal{E}$  generates dense motion tokens through a network consisting of 1D convolutions that are quantized using codes from the codebook, which  $\mathcal{D}$ , also based on 1D convolutions, reconstructs into sequences (see Figure 4). In our model, the encoder takes a sequence of sign poses represented as normalized 2D keypoint coordinates, of length  $M$ , and produces latent vectors  $\hat{z}^{1:L} = \mathcal{E}(m^{1:M})$ , effectively capturing sequences of frames in each latent vector and downsampling a motion sequence  $L = M/l$ , where  $l$  is the downsampling factor. These vectors are then discretized into a set of codebook entries  $z$  through quantization so that each entry  $z_i$  belongs to a learnable codebook  $Z = \{z^i\}_{i=1}^K \subset R^d$ , with  $K$  latent embedding vectors of dimension  $d$ .

$$z_i = Q(\hat{z}^i) := \arg \min_{z_k \in Z} \|\hat{z}^i - z_k\|_2. \quad (1)$$

To reconstruct the sequence the decoder uses these embeddings and outputs a sequence of sign poses of length  $M$ .

Optimization employs reconstruction loss ( $\mathcal{L}_r$ ), which compares the mean squared error between the input and the output of the VQ-VAE, and a commitment loss ( $\mathcal{L}_c$ ) that ensures the encoder com-

mits to an embedding and limits the growth of the embedding space. We also employ other additional techniques for quality enhancement such as replacing the embedding loss ( $\mathcal{L}_e$ ) that minimizes the difference between the encoded sequences and the closest code embeddings, with exponential moving average (EMA) as in [Razavi et al. \(2019\)](#).

## 5. Experiments

For the following experiments we used a codebook size,  $K$ , of 512 and also set the dimension of the embedding vectors,  $d$ , to 512, following [Jiang et al. \(2023\)](#). Based on our analysis of velocity profiles, we used a sequence length,  $M$ , of 30 frames, and for the encoder network, we used a depth of 3 and stride 2 resulting in a downsampling factor,  $l$ , of 7.5 and a latent encoding of length,  $L$ , of 4. This was to ensure that each token in our codebook would correspond to a motion and not only keyframes as in other works such as [Xie et al. \(2022\)](#).

We trained three models on the Swedish Sign Language Dictionary: one using only individual signs (signs model), one using only sentence data (sentences model) and one using all data (mixed model). The data was split 80/10/10 in a train, validation and test set and the same split was used for all models to prevent information leakage.

Test Dataset	Training Dataset		
	Signs	Sentences	Mixed
Signs	0.0067	0.0214	0.0077
Sentences	0.0074	0.0044	0.0074
YouTube	0.0211	0.0146	0.0157

Table 1: Reconstruction loss for models trained on different subsets of the Swedish sign data measured as the mean squared error between the input and the output of the VQ-VAE

As can be seen in Table 1 the models trained on only signs or sentences exhibited better reconstruction for the type of data they were trained on, which was expected. It can also be seen that the reconstruction loss on data from YouTube was lower for the model trained on sentences.

To further investigate how the models learn to represent motion primitives in the codebook, we evaluated the use of codes for the model trained on all the data for 5000 test and training samples from signs and 5000 test samples from sentences respectively. Figure 5 shows that there is a difference in the usage of the codebook and that the distribution over codes for the samples is more similar between signs than between signs and sentences.

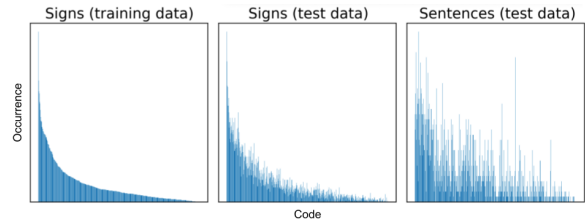


Figure 5: Codebook usage for the model trained on both signs and sentences. The three histograms are sorted horizontally by most used codes for the training data for individual Signs (left).

## 6. Discussion and Conclusion

The results in this paper indicate that it is possible to capture some of the dynamic nature of signing using an unsupervised model such as a VQ-VAE. As is seen in the reconstruction results between the different models and on the different types of data, it is clear that capturing motion primitives more similar to the dynamic of the target data yields better results (see Table 1).

### 6.1. Time Dependence

In its current setup, the VQ-VAE model architecture puts fixed limits on a sequence length, which means data is cut and/or padded to deal with different lengths of motions. The previous works usually set a fixed sequence length based on the domain of the data. For instance, the authors of [Siyao et al. \(2022\)](#) use longer sequence length in their model – 240 frames, compared to co-speech gestures paper [Yazdian et al. \(2022\)](#), who use 30. This editing makes it possible to train a model on data of different lengths.

However, if the same kind of motion primitive is performed with a different velocity, it can change the model’s ability to represent it with the same code. In the domain of signing data, the same signs can also be produced at different speeds, so that one motion primitive is produced within a different number of frames. This is supported by the difference we discovered in the types of codes a mixed model learns from different types of data, indicating that the current architecture needs different codes for different velocities. As a result, there is a limit to the current model’s ability to learn and generalize well over different types of data, even when dealing with the exact same signs. This highlights the need to investigate the possibility to create an unsupervised setup that can capture time-invariant motion primitives for this task.

## 6.2. Generating New Samples

Given the limited amount of annotated sign language data, training an unsupervised model that can be used for a downstream task such as sign language production is of great interest. Even though it is possible to directly sample from the codebook of our model, it yields human-like but nonsensical results. Training a class, or language, guided model for code generation could yield more interesting results but is left as future research.

Additionally, by observing sampled and reconstructed sequence data we identify some limitations of the setup such as a need to improve the finger tracking and also increase the expressiveness of the model. For examples of generated sequences we refer to our project page<sup>3</sup>.

## 7. Acknowledgements

This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, by the Swedish Research Council (VR) proj. 2023-04548.

## 8. Bibliographical References

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.
- Carl Börstell. 2023. Extracting sign language articulation from videos with mediapipe. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20.
- Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. [Motiongpt: Human motion as a foreign language](#).
- Johanna Mesch and Lars Wallin. 2015. Gloss annotations in the swedish sign language corpus. *International Journal of Corpus Linguistics*, 20(1):102–120.
- Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012. [Sign Language Resources in Sweden: Dictionary and Corpus](#). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon [Language Resources and Evaluation Conference (LREC)]*, pages 127–130, Paris. European Language Resources Association (ELRA).
- Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. [Generating diverse high-fidelity images with vq-vae-2](#).
- Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059.
- Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. [Wavenet: A generative model for raw audio](#).
- Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2018. [Neural discrete representation learning](#).
- Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. 2022. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*.
- Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220.
- Payam Jome Yazdian, Mo Chen, and Angelica Lim. 2022. Gesture2vec: Clustering gestures using representation learning methods for co-speech gesture generation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3100–3107. IEEE.
- Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. [Mediapipe hands: On-device real-time hand tracking](#).

<sup>3</sup>[www.speech.kth.se/research/vq-sign](http://www.speech.kth.se/research/vq-sign)



Zhenxing Zhou, Vincent WL Tam, and Edmund Y Lam. 2021. Signbert: a bert-based deep learning framework for continuous sign language recognition. *IEEE Access*, 9:161669–161682.

## **9. Language Resource References**

Svenskt teckenspråkslexikon. 2024. *Swedish Sign Language Dictionary online*. Sign Language Section, Department of Linguistics, Stockholm University.

# Quantitative Analysis of Hand Locations in both Sign Language and Non-linguistic Gesture Videos

Niels Martínez-Guevara\*, Arturo Curiel†

\*Universidad Autonoma de Coahuila, †Independent

\*Centro de Investigación en Matemáticas Aplicadas. Unidad Saltillo. Saltillo, Coahuila, 25280, México.  
niels.martinez@uadec.edu.mx, me@arturocuriel.com

## Abstract

This paper explores whether measurable quantitative linguistic relationships are readily apparent in the use of space of three different Sign Languages (SLs): British Sign Language (BSL), Dutch Sign Language (NGT) and Mexican Sign Language (LSM). To this end, three SL datasets were collected; one for each of the languages of interest. Informative video frames were extracted from the collected datasets, which in turn were automatically processed to detect hand locations. The obtained information was analyzed through statistical methods, and compared against a dataset of non-linguistic gestural communication: the latter, in an effort to observe whether space-use differs between linguistic and non-linguistic gestures. The results show that meaningful gestures—regardless of whether they are deemed linguistic or not—seem to induce a spatial hierarchy around the gesturer, disproportionately favoring certain areas during articulation. SLs in particular seem to exert pressure on those areas to become more efficient, as signers appear to concentrate hand activity over more cohesive regions than non-signers. In addition, these results point towards an indirect relationship between culturally-recognized gestures and their surrounding SLs, showing that there is still work to be done on the exploration of iconicity and its effects on gestural communication.

**Keywords:** Zipf's law, signing space, non-linguistic gestures

## 1. Introduction

Quantitative linguistics is the sub-field of linguistics that studies language through empirical mathematical methods (Best et al., 2017), most of which arise from statistics (Johnson, 2008). Previous work on quantitative linguistics has shown that spoken and written languages fulfill statistical laws that can be asserted as language universals; notably Zipf's law of abbreviation (Bentz and Ferrer-i Cancho, 2016; Linders and Louwerse, 2023) (which states that there is a negative relationship between word length and frequency) and Menzerath-Altmann's law (Eroglu, 2013; Milička, 2014) (which states that larger linguistic structures have shorter constituents and vice versa). Research in the field has also delved into the study of animal communication systems (Ferrer-i Cancho and McCowan, 2009; Heesen et al., 2019; Clink et al., 2020; Huang et al., 2020; Safryghin et al., 2022), and how their rudimentary encoding of meaning produces patterns reminiscent of both laws. However, quantitative linguistic laws have seldom been confirmed in more than a few Sign Languages (SLs) (Malaia et al., 2023); thus, even though SL research has emerged as a compelling area of study for the exploration of statistical linguistic universals, little work has been directed towards the study of quantitative relationships akin to the ones observed in spoken language.

This paper explores the existence of quantitative spatial relationships in three different SLs: British

Sign Language (BSL), Dutch Sign Language (NGT) and Mexican Sign Language (LSM). To this end, four SL datasets were analyzed: three dictionaries and one continuous signing video. Relevant frames were extracted from each collection, which in turn were processed to automatically detect hand locations. Location points were then analyzed with statistical methods, in an effort to discover whether signers assign a strict hierarchy in the signing space consistent with previous observations in quantitative linguistics. The obtained measurements were compared against a dataset of non-linguistic gestural communication videos, so as to explore the differences between linguistic and non-linguistic gestures.

The results show that communicative gestures—whether they are SL or not—seem to induce a spatial hierarchy, disproportionately favoring certain space regions for articulation. SLs in particular seem to exert pressure on those areas to become more efficient, as signers appear to concentrate hand activity over more cohesive regions than non-signers.

The rest of this paper is organized as follows. Section 2 presents some of the existing work in quantitative linguistics for SLs. Section 3 presents the methodology, whereas Section 4 shows the obtained results. Finally, Sections 5 and 6 present the discussion and conclusions, respectively.

## 2. Related work

The volume of existing research in quantitative linguistics strongly implies that SLs must fulfill (at least) the same statistical patterns as spoken languages—even despite their highly iconic nature. However, few works have been directed towards their quantitative exploration.

Among these, [Riedl and Sperling \(1988\)](#) attempted to measure how American Sign Language (ASL) intelligibility is affected depending on changes on the visual signal. The authors filtered the videos of an ASL corpus of isolated signs into different spatiotemporal bands; afterwards, they measured how combining them or adding noise improved (or decreased) intelligibility with Deaf individuals. They found that they could divide isolated signing videos into four high intelligibility bands with enough visual information to discriminate between them; in essence, proving that the discrete nature of language is preserved in SLs regardless of modality.

More recently [Stewart \(2014\)](#) studied how role shifting, sign-type or information status (new vs. given) may affect the duration of ASL signs. The author found that duration (in milliseconds) can be used to distinguish between lexicalized signs and non-conventionalized forms (*i.e.* iconic), pointing towards an underlying meaning-length relationship akin to the law of abbreviation.

Similarly, [Börstell et al. \(2016\)](#) analyzed the relationship between sign duration and frequency in Swedish Sign Language (STS). The authors showed that high-frequency signs in their corpus had shorter durations than low-frequency signs. Also, they showed that signs that act as function words had shorter durations than content signs, once again pointing towards an underlying length-meaning relationship.

[Caselli et al. \(2017\)](#) presented a lexical database of 1000 ASL signs containing information including frequency (as estimated by users), duration, iconicity rating, grammatical class and the signs' phonological properties. Having these measurements enabled the authors to calculate statistical relationships between them; notably, in contrast to previous works, they also took into account sub-lexical features. Their results show that:

- shorter signs were more frequent;
- less iconic signs were more frequent; and,
- the frequencies of individual phonological properties (including location) tended to approximate a power-law distribution.

[Bosworth et al. \(2019\)](#) also analyzed sub-lexical characteristics of ASL, measuring spatiotemporal

properties such as: hand location, hand eccentricity in the visual space, hand motion speed and total traveled distance of the dominant hand. As their predecessors, they also calculated sign duration. The authors found that signers produce *asymmetries* in the visual field (concentrating movement around certain areas). In that regard, their results show that the statistical laws underlying SLs may not only express themselves temporally, but also spatially.

[Fenlon et al. \(2019\)](#) analyzed the difference between linguistic and non-linguistic gestures in SL. The authors compared how pointing signs (with grammatical function) in BSL differed from the pointing gestures produced by non-signing American English speakers. To this end they annotated features such as hand-shape, number of hands, duration and body-contact of the observed pointing instances (in both corpora). Their results show that there is an evolutionary pressure consistent with Zipf's law of abbreviation that makes pointing signs both systematically shorter than pointing gestures, and more stable shape-wise upon production. The authors emphasize that this reduction is expressed along several formational parameters (not only duration) and that it may be related to the high frequency of pointing signs in BSL.

A similar observation was made by [Flaherty et al. \(2023\)](#), regarding the signing space. The authors compared the signing of young and old Nicaraguan Sign Language (ISN) signers using motion tracking technology. Their comparison was based on measuring the size of the 3D space that the signers actually used during production, as well as the average body-wrist distance. The results show that younger signers tended to use less space than older signers, pointing towards a reduction of the signing space consistent with an underlying linguistic optimization model.

## 3. Methodology

The experiments consisted in automatically extracting hand locations from both SL and non-SL gestural videos, so as to explore their respective spatial characteristics. To this end four publicly available SL resources were collected, as well as a non-SL communication dataset.

### 3.1. Datasets

For SL communication three dictionaries were chosen:

- BSL ([Waters, 2003](#)) with 280 signs;
- NGT ([Els van der Kooij, 2003](#)) with 250 signs; and,

- LSM (Alvarez Hidalgo et al., 2009) with 300 signs.

Dictionaries were preferred over continuous signing videos so as to remain fully comparable with the non-SL videos. However, to account for potential changes due to the grammatical use of space, a continuous signing dataset was compiled:

- LSM (continuous) (López-Obrador, 2023), extracted from a publicly available government conference.

Regarding non-SL gestural communication, a video dataset of pantomimes, emblems<sup>1</sup> and meaningless gestures was chosen (Lingnau, 2018), containing the following video distribution (Agostini et al., 2019):

- Emblems (103 videos);
- Pantomimes (90 videos); and,
- Meaningless (77 videos).

Notably, the gestures represented in the dataset were rated on how meaningful they were deemed by American and Italian raters; with Pantomimes showing a higher consensus on their apparent meaning than Emblems. These differences between them may be important for comparison against SL signs, as it means that some gestures may share an iconic “root” with some signs (particularly Pantomimes). Moreover, lower consensus on the meaning of Emblems might also point towards cultural differences that may affect the creation of meaningful communication symbols—which, in turn, could potentially have a measurable effect on SLs.

From this dataset, only Emblems and Pantomimes were analyzed; the ambiguous nature of the Meaningless videos made them difficult to interpret when compared against signing videos. Thus, in the end, a total of six collections were considered: three SL dictionaries, one continuous signing video and two non-SL video datasets.

### 3.2. Frame extraction

For this study only a subset of informative video frames were considered from each dataset. Mainly, in an effort to avoid over-representation of space regions across collections, which could be biased by frame rate differences or changes in signing speed. This strategy also served to reduce the computational overhead of the analysis.

Thus, all relevant video frames were automatically extracted from each of the six aforementioned collections. The extraction process followed the algorithm proposed by Martinez-Guevara et al.

<sup>1</sup>Gestures with a culturally agreed-upon meaning.

(2023), based on finding stable *fixed postures*: video frames with minimal change with respect to a context window, as given by the Structural Similarity Index (SSIM) (Wang et al., 2004). The authors showed that the extracted frames are relevant in the sense that they contain enough information for native signers to still understand the utterance if presented with those frames alone; *i.e.* they contain enough information to preserve the message.

Table 1 shows the quantity of fixed postures extracted from each collection. Note that fixed postures were obtained from at most 90 random signs per dataset, so as to remain consistent with the number of gestures available in the Pantomimes dataset.

Dataset	NO. OF FIXED POSTURES	NO. OF GESTURES OR SIGNS
BSL	137	44
NGT	136	44
LSM	272	86
LSM (continuous)	280	≈90
Emblems	133	90
Pantomimes	148	90

Table 1: Number of fixed postures (frames) extracted from each dataset.

As implied by Table 1, during the extraction process some signs had to be discarded due to issues arising from the extraction script: namely, with the oldest datasets (BSL and NGT) the algorithm had trouble distinguishing between similar frames. In part, because of image noise. The same happened with four out of the 90 LSM signs; however, it was far less common as the LSM videos were of decidedly better quality than the others (higher resolution and less noise). For the continuous LSM collection, frame extraction was artificially capped to 280 frames, assuming it would correspond to approximately 90 signs (following the values obtained from the dictionary videos).

The idea of the extraction is roughly based on the phonetic/phonological models proposed by Liddell and Johnson (1989); Johnson and Liddell (2011). Fixed postures would approximate *Holds* in the original phonological model, or *postural segments* in the phonetic framework.

### 3.3. Hand location extraction

For the analysis, the extracted fixed postures were labeled with OpenPose (Simon et al., 2017; Cao et al., 2019): a body location detection toolkit capable of detecting the 2D positions of up to 135 keypoints. Figure 1 shows the toolkit’s output on a



single frame.



Figure 1: Keypoint posture detection with OpenPose.

For each fixed posture, only two points of interest were considered: the left and right hand locations, which were assumed to be indicative of place of articulation. In that regard, both keypoints would be able to show the entire extent of the signing space; notably, enabling the study of regions with out-sized importance for communication—those concentrating the most activity across multiple fixed postures. Figure 2 shows a scatter plot with all the points extracted from the continuous LSM dataset.

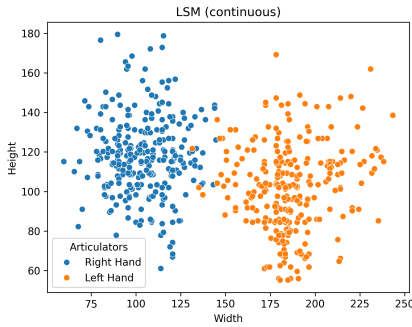


Figure 2: Points extracted from the continuous LSM dataset.

The obtained points were normalized prior to their analysis using the following formulae:

$$\hat{x}_i^k = \frac{x_i^k}{W_i} \quad (1)$$

$$\hat{y}_i^k = \frac{y_i^k}{H_i} \quad (2)$$

where:

- $(x_i^k, y_i^k)$  denotes the  $k$ -th point in collection  $i$ , in pixels.
- $H_i$  is the pixel height of the videos contained in collection  $i$ .
- $W_i$  is the pixel width of the videos contained in collection  $i$ .

The resulting  $(\hat{x}_i^k, \hat{y}_i^k)$  normalized points were defined in the interval  $[0, 1]$ , regardless of the source.

This enabled the direct comparison of space regions between datasets, using both classical Euclidean metrics and clustering evaluation scores.

Note that the normalization procedure didn't consider intrinsic features such as body size of the signer or proximity to the camera. However this shouldn't pose problems for the analysis, save for the comparison of point dispersion across datasets through standard deviation. The results are shown in Section 4.1.

### 3.4. Location density analysis

The normalized point clouds induced by the previously described procedure enabled the approximation of a location Probability Density Function (PDF) for each dataset, showing where activity tended to concentrate across relevant frames. The approximation was performed through Kernel Density Estimation (KDE) (Sheather, 2004), using the Python implementation included with the Scikit-learn library (Pedregosa et al., 2011). Scott's rule was used to determine the optimal bandwidth. A visual depiction of the obtained densities can be observed in Figure 3.

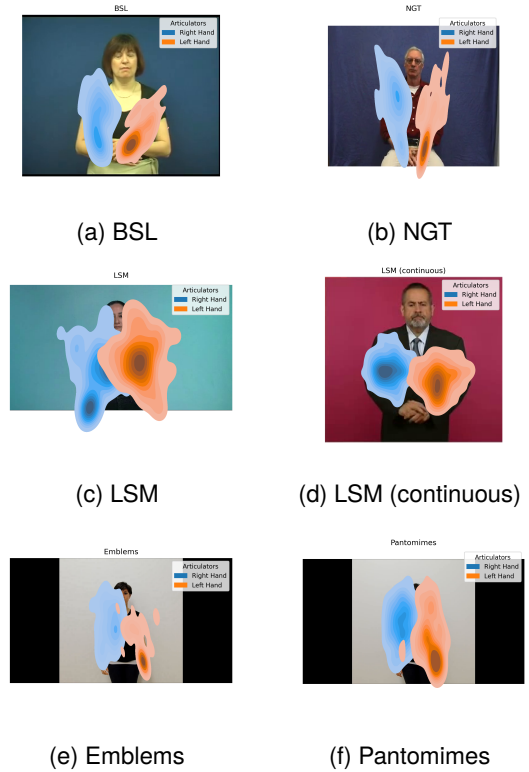


Figure 3: Location density maps for the six analyzed datasets.

In the Figure, darker densities imply higher activity: this is, there is a stronger probability of a hand being active in said region, for any given fixed posture. Regions with no color overlap denote zero

(or nearly zero) probability of including a hand.

The obtained PDFs were sampled and compared against other probability distributions in two distinct cases:

- To determine their goodness-of-fit against a power law distribution, so as to show whether locations are *Zipfian* in nature.
- To compare whether the use of space changes between SL and non-SL gestures.

The comparisons were performed by way of the Kolmogorov-Smirnov test. The results are shown in Section 4.2.

## 4. Results

In order to determine whether the use of space changes depending on the type of dataset, two kinds of measurements were taken from the extracted data:

- **Dispersion and separation measures:** to observe if there is a measurable *cohesion* regarding the spatial distribution of the hands across multiple signs (*i.e.* if the signing space tends to “shrink”).
- **Hypothesis testing:** to confirm whether there is a hierarchical relationship between space regions, as given by location densities (*i.e.* if signers will disproportionately “prefer” some regions above others).

### 4.1. Dispersion and separation measures

Dispersion was measured in terms of the Euclidean distance between the normalized points described in Section 3.3. For each dataset, the pairwise distances between all the extracted points were calculated. Table 2 shows both the mean distance and the obtained standard deviation for each of the six collections.

Dataset	RIGHT		LEFT	
	$\mu$	$\sigma$	$\mu$	$\sigma$
BSL	0.155	0.095	0.109	0.085
NGT	0.187	0.135	0.156	0.174
LSM	0.231	0.155	0.206	0.134
LSM (cont.)	0.118	0.063	0.131	0.071
Emblems	0.161	0.108	0.098	0.124
Pantomimes	0.175	0.110	0.196	0.155

Table 2: Mean ( $\mu$ ) Euclidean distances and their standard deviation ( $\sigma$ ) for all datasets.

Note that the measurements in Table 2 are separated by hand: as hands are able to act with relative

independence with respect to one another, it is expected that they’d have their own preferred regions of activity. Thus, any measurable pressure or spatial hierarchy should be independently observable in at least one of the hands.

Separation between the hands’ regions was measured by way of two intrinsic clustering evaluation metrics: the Silhouette Coefficient (Rousseeuw, 1987) and the SDbw validity index (Halkidi and Vazirgiannis, 2001).

The Silhouette Coefficient measures the difference between the average intra-cluster distance (*i.e.* calculated between the points in the group) and the average inter-cluster distance (*i.e.* calculated between the points outside the group). It is defined on the interval  $[-1, 1]$ , where -1 denotes poor cluster separation and 1 denotes perfect cluster separation.

The SDbw validity index measures the difference between the average intra-cluster distance and the average inter-cluster point density (*i.e.* the distance between the cluster centroids). The resulting value is higher than zero, with **lower** values denoting better cluster separation. Table 3 shows the calculated scores for each dataset.

Dataset	SILHOUETTE	SDBW
BSL	0.481	0.717
NGT	0.473	0.741
LSM	0.271	0.855
LSM (cont.)	0.580	0.569
Emblems	0.512	0.670
Pantomimes	0.285	1.352

Table 3: Silhouette coefficient and SDbw validity index for all point clouds.

Together, these results show how cohesive the use of space is in the tested datasets. However, they don’t show whether there might be a spatial hierarchy between specific regions, as indicated by hand activity. For the latter, measurements over location probability densities (rather than individual points) had to be considered, as presented in the next section.

### 4.2. Hypothesis testing

For these experiments, the estimated PDFs described in Section 3.4 were sampled and compared against a power law distribution. The comparison was performed by way of a two-sided Kolmogorov-Smirnov test, using the Scipy library (Virtanen et al., 2020). The results are shown in Table 4.

As with the clustering experiments, hands were measured separately. Note that with a  $p = 0.05$  significance level, the null-hypothesis (samples come from the same distribution) **cannot** be rejected for any case.

KS TEST				
Dataset	RIGHT HAND		LEFT HAND	
	$D$	$p$	$D$	$p$
BSL	0.106	0.19	0.062	0.81
NGT	0.113	0.14	0.058	0.86
LSM	0.064	0.78	0.095	0.30
LSM (cont.)	0.067	0.74	0.075	0.59
Emblems	0.111	0.15	0.054	0.91
Pantomimes	0.106	0.19	0.087	0.41

Table 4: Kolmogorov-Smirnov test results comparing location densities against a power law distribution.

To complement these results, Figure 4 shows six plots describing how a power law distribution fits the location density data. Only the values for the dominant hand are displayed.

In Figure 4, note that the following  $\alpha$  parameters were estimated for the depicted power laws:

- BSL  $\alpha = 2.81$
- NGT  $\alpha = 2.44$
- LSM  $\alpha = 2.81$
- LSM (continuous)  $\alpha = 2.53$
- Emblems  $\alpha = 2.02$
- Pantomimes  $\alpha = 1.97$

Finally, Table 5 shows the comparison of location density distributions between SL and non-SL datasets. As before, the comparison was performed by way of a two-sided Kolmogorov-Smirnov test.

Note that  $D$  denotes the maximum absolute difference between the two tested distributions: this is, a higher  $D$  indicates that they are very different from one another, whereas a lower  $D$  indicates that they are very similar. The results are discussed in the next section.

## 5. Discussion

In general, the results support the notion that conveying meaning puts pressure on the use of space during gestural communication, regardless of whether it is SL or not.

For instance, looking at the Euclidian distance statistics, it can be observed that both signers and non-signers tend to concentrate movement around certain regions: the results in Table 2 show that the average distance between locations (for all cases) tends to be below 20% of the available space, with a systematically lower-than-the-mean standard deviation pointing towards low dispersion. Regarding the dominant hand (the right hand in all collections), the

KS TEST		
Dataset	RIGHT HANDS	
	$D$	$p$
Emb. - BSL	0.279	$3.79 \times 10^{-5}$
Emb. - NGT	0.323	$9.50 \times 10^{-7}$
Emb. - LSM	0.705	$1.76 \times 10^{-43}$
Emb. - LSM (cont.)	1.0	$2.01 \times 10^{-111}$
Pant. - BSL	0.324	$4.02 \times 10^{-7}$
Pant. - NGT	0.594	$7.38 \times 10^{-24}$
Pant. - LSM	0.647	$1.68 \times 10^{-38}$
Pant. - LSM (cont.)	1.0	$1.61 \times 10^{-118}$
Pant. - Emb.	0.479	$3.65 \times 10^{-15}$
Dataset	LEFT HANDS	
	$D$	$p$
Emb. - BSL	0.698	$3.95 \times 10^{-32}$
Emb. - NGT	0.812	$4.98 \times 10^{-45}$
Emb. - LSM	0.838	$2.77 \times 10^{-65}$
Emb. - LSM (cont.)	1.0	$2.01 \times 10^{-111}$
Pant. - BSL	0.788	$2.30 \times 10^{-44}$
Pant. - NGT	0.396	$1.64 \times 10^{-10}$
Pant. - LSM	0.849	$5.37 \times 10^{-72}$
Pant. - LSM (cont.)	1.0	$1.61 \times 10^{-118}$
Pant. - Emb.	0.789	$5.32 \times 10^{-44}$

Table 5: Kolmogorov-Smirnov test results comparing the location density distributions between SL and non-SL datasets.

continuous signing video was the one that covered the shortest distance. This was to be expected: lexicons are intended to show signs in a clear, systematic, manner, whereas continuous signing intends to convey a concrete message—implying that communication has to be more efficient, thus limiting the breadth of movement to its minimal expression. As such, it is not surprising that continuous signing had the lowest standard deviation of all collections. However, as implied before, this could also be an effect of camera positioning or the fact that the signer is aware of the space limitations he has—taking into account the fact that the continuous signing example comes from an interpretation task.

Similarly, the clustering results from Table 3 show that the continuous LSM dataset provided a stronger definition of hand regions, whereas pantomimes tended to be remarkably less stable than both SLs and emblems alike. A notable exception is the LSM lexicon, which shows a lower Silhouette score than pantomimes; however, when accounting for point density, its cluster definition became closer to the remaining lexicons rather than to pantomimes or emblems. Essentially, implying that there is a more systematic use of space in the former that is not well established in the latter. This can be partially seen in Figure 3 as well, where it can be observed that the use of space in the Pantomimes dataset tends to be less focused than in

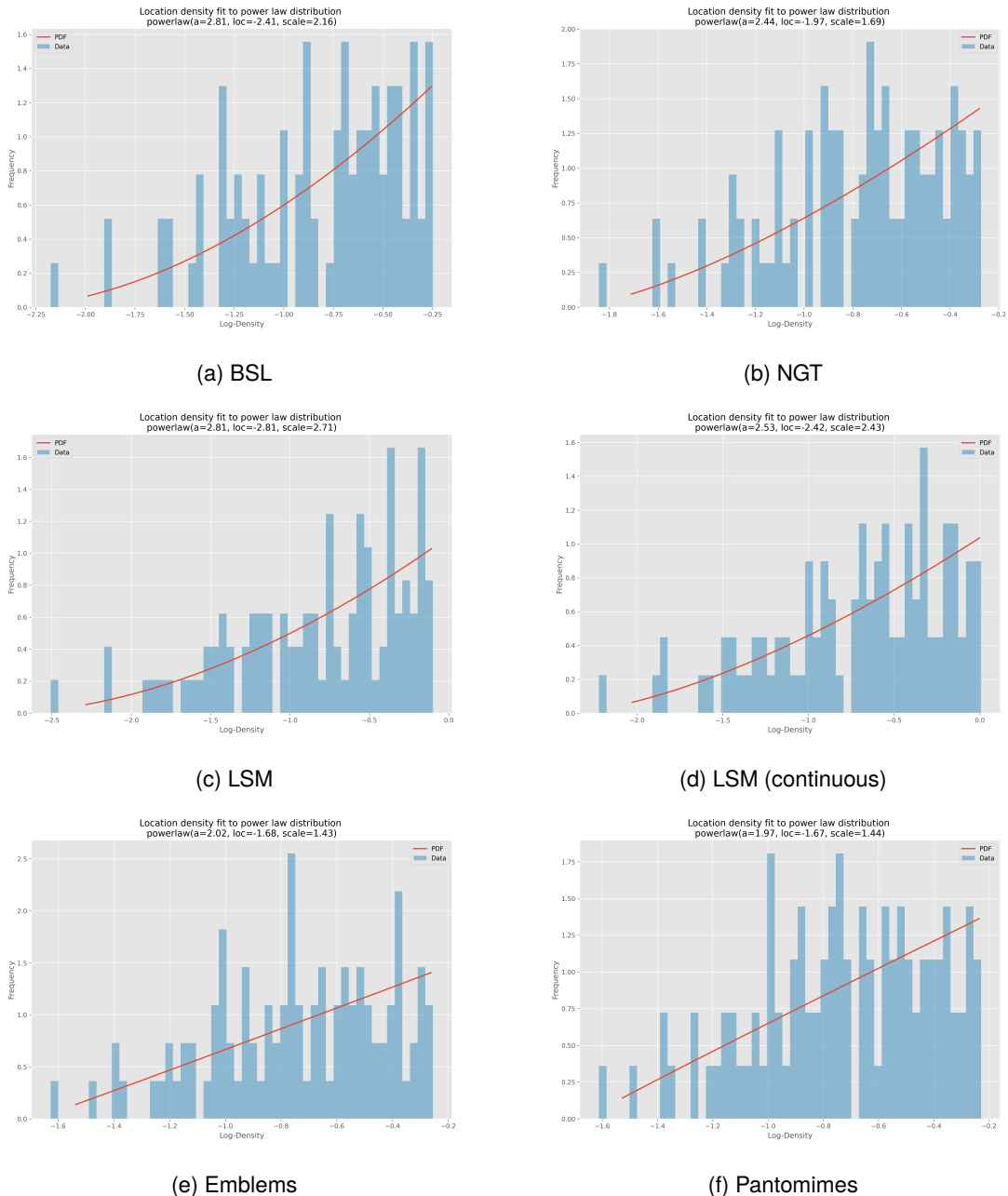


Figure 4: Dominant hand location log-density data fitted to a power law distribution.

the SL lexicons.

Regarding the location density results, Table 4 shows that not only do signers limit their movement to specific space regions, but they do so in a *Zipfian* manner: some regions are exponentially more active than others. The six collections showed tendency to this phenomenon, consistent with previous observations on the effects of meaning on natural communication systems. Nevertheless, when compared to SL datasets, pantomimes and emblems showed a marginally lower growth on their calculated power law distribution; this may indicate that a spatial hierarchy already exists in non-linguistic communication, but it is less strict than the one

induced by SLs.

Finally, the direct comparison between SL and non-SL datasets shows that, even though the density distributions are decidedly different from one another, they are close enough to warrant further explanation—at least, with respect to BSL and NGT. For instance, Table 5 shows that space-use in the Emblems videos is surprisingly similar to BSL; this could be due to the fact that the former dataset was created considering cultural gestures in mind, which could very well be represented in BSL. Thus, there could be an underlying relationship not readily apparent between the two: they could share the same iconic DNA due to cultural



proximity. Nonetheless, there is not enough information in the selected datasets to confirm the existence of such a relationship.

In the end, the obtained results seem to show the existence of a spatial hierarchy linked to the act of conveying meaning. However, the scale of the performed experiments was too limited: only one signer-gesturer was present in each of the six collections. Furthermore, differences between digital media (e.g. image size, frame-rate, etc); the kind of dataset; noise introduced by OpenPose; or the accidental extraction of non-relevant frames may be acting as sources of bias that are difficult to interpret within the chosen framework of analysis. Ideally, an homogeneous parallel corpus would be better suited to explore the existence of quantitative linguistic laws. Thus, further experiments are required—on a larger scale—to confirm the presented results.

## 6. Conclusions

The quantitative exploration of SLs constitutes one additional step towards improving our understanding of the diversity of human language. The present study contributes to these efforts by showing that gestural communication seems to induce a measurable spatial hierarchy, that follows a probability distribution related to Zipf's law. Moreover the obtained results show that, contrary to non-linguistic gestures, SLs tend to systematize the use of space to optimize information exchange. Nonetheless, future research is needed to confirm these observations in larger, homogeneous corpora. Additionally, some results indicate that it may be worth it to explore the connection between culturally-recognized gestures and their surrounding SLs, as the articulation of the latter may be disproportionately influenced by the former. In that regard, understanding how both processes connect may also shed light on how iconicity influences SL morphology, leading to sign formation.

## 7. References

Beatrice Agostini, Liuba Papeo, Cristina-Ioana Galusca, and Angelika Lingnau. 2019. [A norming study of high-quality video clips of pantomimes, emblems, and meaningless gestures](#). *Behavior Research Methods*, 51(6):2817–2826.

Christian Bentz and Ramon Ferrer-i Cancho. 2016. [Zipf's law of abbreviation as a language universal](#). Leiden, The Netherlands.

K.H. Best, O. Rottmann, and RAM-Verlag. 2017.

[Quantitative Linguistics, an Invitation](#). Studies in quantitative linguistics. RAM-Verlag.

Rain G. Bosworth, Charles E. Wright, and Karen R. Dobkins. 2019. [Analysis of the visual spatiotemporal properties of American Sign Language](#). *Vision Research*, 164:34–43.

Carl Börstell, Thomas Hörberg, and Robert Östling. 2016. [Distribution and duration of signs and parts of speech in Swedish Sign Language](#). *Sign Language & Linguistics*, 19(2):143–196.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. [Openpose: Realtime multi-person 2d pose estimation using part affinity fields](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Naomi K. Caselli, Zed Sevcikova Sehyr, Ariel M. Cohen-Goldberg, and Karen Emmorey. 2017. [ASL-LEX: A lexical database of American Sign Language](#). *Behavior Research Methods*, 49(2):784–801.

Dena J. Clink, Abdul Hamid Ahmad, and Holger Klinck. 2020. [Brevity is not a universal in animal communication: evidence for compression depends on the unit of analysis in small ape vocalizations](#). *Royal Society Open Science*, 7(4):200151. Publisher: Royal Society.

Sertac Eroglu. 2013. [Menzerath–Altmann law for distinct word distribution analysis in a large text](#). *Physica A: Statistical Mechanics and its Applications*, 392(12):2775–2780.

Jordan Fenlon, Kensy Cooperrider, Jon Keane, Diane Brentari, and Susan Goldin-Meadow. 2019. [Comparing sign language and gesture: Insights from pointing](#). *Glossa*, 4(1).

Ramon Ferrer-i Cancho and Brenda McCowan. 2009. [A Law of Word Meaning in Dolphin Whistle Types](#). *Entropy*, 11(4):688–701.

Molly Flaherty, Asha Sato, and Simon Kirby. 2023. [Documenting a Reduction in Signing Space in Nicaraguan Sign Language Using Depth and Motion Capture](#). *Cognitive Science*, 47(4):e13277.

M. Halkidi and M. Vazirgiannis. 2001. [Clustering validity assessment: finding the optimal partitioning of a data set](#). In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 187–194.

Raphaela Heesen, Catherine Hobaiter, Ramon Ferrer-i Cancho, and Stuart Semple. 2019. [Linguistic laws in chimpanzee gestural communication](#). *Proceedings of the Royal Society B: Biological Sciences*, 286(1896):20182900. Publisher: Royal Society.

- Mingpan Huang, Haigang Ma, Changyong Ma, Paul A. Garber, and Pengfei Fan. 2020. [Male gibbon loud morning calls conform to Zipf's law of brevity and Menzerath's law: insights into the origin of human language.](#) *Animal Behaviour*, 160:145–155.
- Keith Johnson. 2008. *Quantitative Methods In Linguistics*. John Wiley & Sons. Google-Books-ID: M5uGEAAAQBAJ.
- Robert E Johnson and Scott K Liddell. 2011. A segmental framework for representing signs phonetically. *Sign Language Studies*, 11(3):408–463.
- Scott K Liddell and Robert E Johnson. 1989. American sign language: The phonological base. *Sign language studies*, 64(1):195–277.
- Guido M. Linders and Max M. Louwerse. 2023. [Zipf's law revisited: Spoken dialog, linguistic units, parameters, and the principle of least effort.](#) *Psychonomic Bulletin & Review*, 30(1):77–101.
- Evie A. Malaia, Joshua D. Borneman, Emre Kurtoglu, Sevgi Z. Gurbuz, Darrin Griffin, Chris Crawford, and Ali C. Gurbuz. 2023. [Complexity in sign languages.](#) *Linguistics Vanguard*, 9(s1):121–131. Publisher: De Gruyter Mouton.
- Niels Martinez-Guevara, Jose-Rafael Rojano-Caceres, and Arturo Curiel. 2023. Unsupervised extraction of phonetic units in sign language videos for natural language processing. *Universal Access in the Information Society*, 22(4):1143–1151.
- Jiří Milička. 2014. [Menzerath's Law: The Whole is Greater than the Sum of its Parts.](#) *Journal of Quantitative Linguistics*, 21(2):85–99. Publisher: Routledge \_eprint: <https://doi.org/10.1080/09296174.2014.882187>.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Thomas R. Riedl and George Sperling. 1988. [Spatial-frequency bands in complex visual stimuli: American Sign Language.](#) *JOSA A*, 5(4):606–616. Publisher: Optica Publishing Group.
- Peter J. Rousseeuw. 1987. [Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.](#) *Journal of Computational and Applied Mathematics*, 20:53–65.
- Alexandra Safryghin, Catharine Cross, Brittany Falon, Raphaela Heesen, Ramon Ferrer-i Cancho, and Catherine Hobaiter. 2022. [Variable expression of linguistic laws in ape gesture: a case study from chimpanzee sexual solicitation.](#) *Royal Society Open Science*, 9(11):220849. Publisher: Royal Society.
- Simon J. Sheather. 2004. [Density estimation.](#) *Statistical Science*, 19(4):588–597.
- Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*.
- Jesse Stewart. 2014. [A quantitative analysis of sign lengthening in American Sign Language.](#) *Sign Language & Linguistics*, 17(1):82–101. Publisher: John Benjamins.
- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. [SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.](#) *Nature Methods*, 17:261–272.
- Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. [Image quality assessment: from error visibility to structural similarity.](#) *IEEE Transactions on Image Processing*, 13(4):600–612.

## 8. Language Resource References

- Alvarez Hidalgo, A. and Acosta Arrellano, A. and Moctezuma Contreras, C. and Sanabria Ramos, E. and Maya Ortega, E. and Álvarez Hidalgo, G. and Márquez Vaca, M. and Sanabria Ramos, M. and Romero Rojas, N. 2009. *Dielseme 2 Diccionario de Lengua de Señas Mexicana*. Secretaría de Educación Pública. <http://campusdee.ddns.net/publicacionesdee.aspx>.
- Els van der Kooij, Annika Nonhebel & Wim Emmerik. 2003. *ECHO NGT lexicon, Male signer*. "ECHO", The Language Archive.

PID <https://hdl.handle.net/1839/00-0000-0000-0008-1763-3>.

Lingnau, Angelika. 2018. *Pantomimes, emblems, and meaningless gestures*. Royal Holloway, University of London. PID <https://doi.org/10.17637/rh.c.4219988>.

López-Obrador, Andres-Manuel. 2023. *#ConferenciaPresidente | Jueves 19 de octubre de 2023*. Presidencia de la República. [https://www.youtube.com/watch?v=kU4VYK\\_RoRg](https://www.youtube.com/watch?v=kU4VYK_RoRg).

Dafydd Waters. 2003. *BSL Lexicon CN*. "ECHO", The Language Archive. PID <https://hdl.handle.net/1839/00-0000-0000-0008-1768-5>.

# Formal Representation of Interrogation in French Sign Language

Emmanuella Martinod<sup>id</sup>, Michael Filhol<sup>id</sup>

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)  
507, rue du Belvédère, 91400, Orsay, France  
{emmanuella.martinod, michael.filhol}@lisn.upsaclay.fr

## Abstract

This paper concerns the marking of interrogation in French Sign Language (LSF). Early work on Sign Languages (SLs) underlined the role of non-manual elements in the production of interrogatives. Studies often point to the role of eyebrows depending on the type of question: eyebrows would usually be raised for the production of yes/no questions, while they would be lowered for other types of questions. For LSF, previous studies seem to validate this contrast. We tested this thoroughly in the framework of AZee, a formal approach to SL modeling based on the identification of linguistic associations between forms and identified meanings, called *production rules*. We present our methodology to extract AZee *production rules*, consisting of data searches alternating form and meaning criteria gradually converging to strong associations, ultimately leading to production rules. Our results (i) show no link between raised or lowered eyebrows and a specific type of question, (ii) highlight instead the role of another non-manual marker: the advancement of the chin. However, since eyebrows remain frequently involved in the analyzed questions (all types included), we intend to further focus on the potential role of the signer's expectations while formulating his request.

**Keywords:** Sign language, Formal representation, Interrogation, Non-manual markers, LSF, SL Synthesis, AZee

## 1. Introduction

This article deals with a specific problem: interrogatives in French Sign Language (LSF), hitherto unaddressed through a formal approach. However, this phenomenon is essential if one wants to generate dialogues in Sign Language (SL), particularly in the case of signing avatars.

Current approaches to describe SLs formally are often elaborated from spoken languages, which are linear systems (see Hadjadj, Filhol, and Braf-fort (2018) for a review of existing systems). This may pose some fundamental problems since SLs are multi-linear visual-gestural languages. In contrast, the AZee model aims at integrating all the forms and phenomena observable in SL (Filhol, 2008, 2021). It is a corpus-based approach that defines systematic links between observed forms and interpreted meanings. It allows a formal representation of SL utterances. Our general goal is to extend the LSF coverage with AZee.

The following section (section 2) gives an overview of claims from previous studies on the topic of interrogatives for SL, and more specifically for LSF. We briefly present the basics of AZee approach and the methodology to enrich its system in section 3, after what we introduce the data we analyzed, and detail the application of the methodology on the data (section 4). Then, we show how our results confirm previous claims in literature and generate a way to cover interrogatives with AZee (section 5). Finally, we discuss the contribution of this work (section 6) and we propose some direction for future studies (section 7).

## 2. Interrogatives in Sign Languages

The relevant literature underlines the role of non-manual elements in the production of interrogatives in SLs ((Neidle et al., 2000) for American SL, ASL; (Coerts, 1990; Klomp, 2021) for Dutch SL, NGT; (Sutton-Spence and Woll, 1999) for British SL, BSL; or (Dubuisson et al., 1991) for Quebec SL, LSQ; see also (Cecchetto, 2012) for a review of previous work on several SLs). Most of these studies establish the role of different non-manual markers depending on the type of question. In these studies, the eyebrows seem to be raised and the head in a forward position for the production of so-called closed questions,<sup>1</sup> while the eyebrows would be lowered for the production of so-called open questions.<sup>2</sup>

However, since the beginning of the interest for this subject, some authors have pointed out the complexity of this phenomenon. Firstly, sometimes this dichotomy does not always seem so obvious (Baker and Cokely, 1980; Dubuisson et al., 1991). Secondly, non-manual elements can combine with other markers that have nothing to do with questioning, for instance, emotions (Weast, 2008, 2011; de Vos et al., 2009). Additionally, less studied SLs could display a slightly less marked pattern (Zeshan, 2004; Cañas Peña, 2019).

The only recent publication dealing with LSF,

<sup>1</sup> Closed questions: questions to be answered with "yes" or "no" (e.g. in English, "Is he coming tonight?").

<sup>2</sup> Open questions: questions that can't be answered with a simple "yes" or "no" (e.g. "What is your name?").



(Sallandre et al., 2021), is based on a previous grammar of LSF that is widely used in teaching (Moody, 1983). It tends to validate the formal opposition between closed-ended and open-ended questions.

Finally, when it comes to SL avatar animation, it is often based on SL linguistics, which provides organization rules for animation data. Thus, the distinction between types of questions seems to be used also in this domain (McDonald et al., 2017).

### 3. The AZee approach

Since we chose AZee approach, here is a summarized presentation of its main principles. Then, we explain the corpus-based methodology to enrich the existing AZee system.

#### 3.1. Production rules

AZee is a formal approach to SL modeling. It depends on identified linguistic associations between observable forms (i.e. timed body articulations) and identified meanings, for instance “pretty, beautiful”. These associations are called *production rules*, and are generally given a name. For instance in LSF, production rule `pretty` associates the meaning “pretty, beautiful” with the form given in Figure 1.

Production rules can be parameterized with named arguments, which can be mandatory or optional (Hadjadj et al., 2018). For instance, the rule `pretty` has no argument. In contrast, the rule `inter-subjectivity` which supports meaning “everybody agrees on *sig*” has one argument named *sig*, which represents the object of agreement. The associated form is a lip pout produced over the form of *sig*.<sup>3</sup>

Combining our two example rules, the following expression composes the meaning: “everybody agrees that [it is] beautiful”:

```
:inter-subjectivity
  'sig
  :pretty
```

Combining the associated forms results in fine detail synthesis and articulation synchronisation (Filhol and McDonald, 2018, 2020). Such *discourse expressions* can build up to arbitrary size, reflecting both the signed forms to be produced and the meaning to be interpreted from them.

The set of all production rules found for a given SL constitutes what is called the *AZee production set* for that language. The next section explains how these rules are extracted from corpus data, a

<sup>3</sup> It is worth noting that this form contains only the necessary and sufficient elements associated with the meaning “everybody agrees on *sig*”.

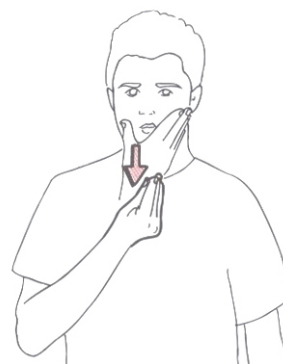


Figure 1: Form for “pretty, beautiful” in LSF

methodology which we will be applying in section 4.

#### 3.2. Rule extraction methodology

Production rules only come from SL data. It is an essential point that makes it a rigorous corpus-based approach. A precise methodology exists to extract AZee production rules from data (Hadjadj et al., 2018). It consists in data searches alternating form and meaning criteria, gradually converging to strong associations ultimately leading to production rules. Form observations are done on videos with the naked eye, so it is the case in the work reported here, although additional software measurements would be possible for more accurate data, in particular for better analysis of dynamics. Meaning interpretation, though, is always assumed to be performed by a human in the process, which is also the case here.

We explain the steps of the process below, as we will be applying it later in section 4:

1. start with an arbitrary form or meaning criterion  $C$  to explore;
2. locate and list all occurrences of  $C$  in a selected SL corpus, and let  $N_{occ}$  be the number of occurrences;
3. for each occurrence of  $C$  listed, add description elements:
  - of interpretation if  $C$  is a form criterion;
  - of observed form if  $C$  is a meaning;
4. identify groups of at least two occurrences with identical description elements, and let  $N_{out}$  be the number of occurrences not included in any group;
5. if all of the following conditions are satisfied:
  - $C$  is a meaning criterion;
  - a unique group was identified in step 4;

- $N_{out}$  is below a threshold, e.g. 15% of  $N_{occ}$ ;

then the form elements defining the unique group  $C.1$  can be considered invariant, and we define a new production rule associating  $C$  with the invariant form, and this iteration stops;

6. if this iteration has not stopped, for each group identified in step 4 defined by semantic or form feature  $f$ :
  - if  $C$  is a meaning that can already be expressed using known production rules justifying form  $f$  or, conversely,  $f$  is a meaning that can already be expressed using known production rules producing form  $C$ , then no new rule is to be found, nor any new search to be fired;
  - otherwise, recursively apply this methodology with a new iteration starting with criterion  $f$ —note that this new search must apply to the whole corpus again, not be restricted to the occurrences defining this group.

## 4. Applying the methodology

We applied the methodology presented in the previous section (section 3) on a data set of LSF, presented first. Then, after refining the notion of “interrogative” for a solid starting criterion, we detail the iterations and the resulting numbers.

### 4.1. Data

The LSF video data we analyzed come from *Dicta-Sign* corpora (Belissen et al., 2020), (Hanke et al., 2010). These are semi-elicited dialogues, likely to contain interrogatives, and already translated into written French. 12 videos (total duration: 1 hour 21 minutes and 47 seconds) produced by 8 dyads of signers were examined.

### 4.2. Starting criterion

The first step to apply the methodology is to choose a starting criterion, of either meaning or form. In this study, being interested in interrogative utterances, we thought of the following meaning criterion: “a question is asked by the signer”. But we faced the issue of determining what was indeed meant by “question” here.

First, we had to exclude what is often called “rhetorical questions” or “question-answer pairs” (Herrmann et al., 2019), which are frequent in LSF, as in other SLs. Their aim is to keep the interlocutor’s attention before introducing new information, but with no real interrogative meaning (e.g.: ‘DATE

WHEN JANUARY’ in LSF, whose sole purpose is to give an information about a date, here January, and not to question the addressee).

Secondly, although available, relying on the French translation of the corpus did not seem relevant here. Indeed, in written French, interrogatives are identified by a question mark. In LSF, the sign “interrogation mark”, tracing the shape of this punctuation sign in the signing space, can be used when asking something to someone but it is far from compulsory. In fact, any request of information in SL could be translated by a French sentence with or without interrogative mark (e.g.: “What is your name?” vs. “Give me your name, please.”).

We therefore clarified our starting criterion as follows. We will be calling this criterion “**IR**” (information request) henceforth:

- the signer is requesting information or confirmation from the addressee;
- the signer does not know the information, but expects the addressee to know;
- the signer expects the addressee to provide it immediately and will wait for it before proceeding.

### 4.3. Running the iterations

We applied the methodology outlined in section 3, starting with our newly defined meaning criterion, IR. An overview of the whole process is given in Figure 2.

The first step is to identify and list all occurrences of criterion IR (meaning criterion of a request of information) in the selected corpus. We found 182 occurrences. For each occurrence of IR, we then indicate elements of form since IR is a meaning criterion. In this case we observed mixes of various form features such as the advancement of the chin (which we will note “AC” henceforth), and posture holds at the end of the production (“H”). We also chose to document eyebrow activity to test the commonly admitted proposition about questions: “RE” for raised eyebrows, “LE” when lowered. Two groups emerged depending mostly on eyebrow activity, as summarized below.  $N_{out} = 25$  covers the entries that fall in neither of the groups. 15 of these entries show no advanced chin or no final hold. The other 10 do show both, but contain no eyebrow activity.

#### Iteration IR (meaning)

- $N_{occ} = 182$
- Groups found:
  1. **AC + RE + H** (96 entries)

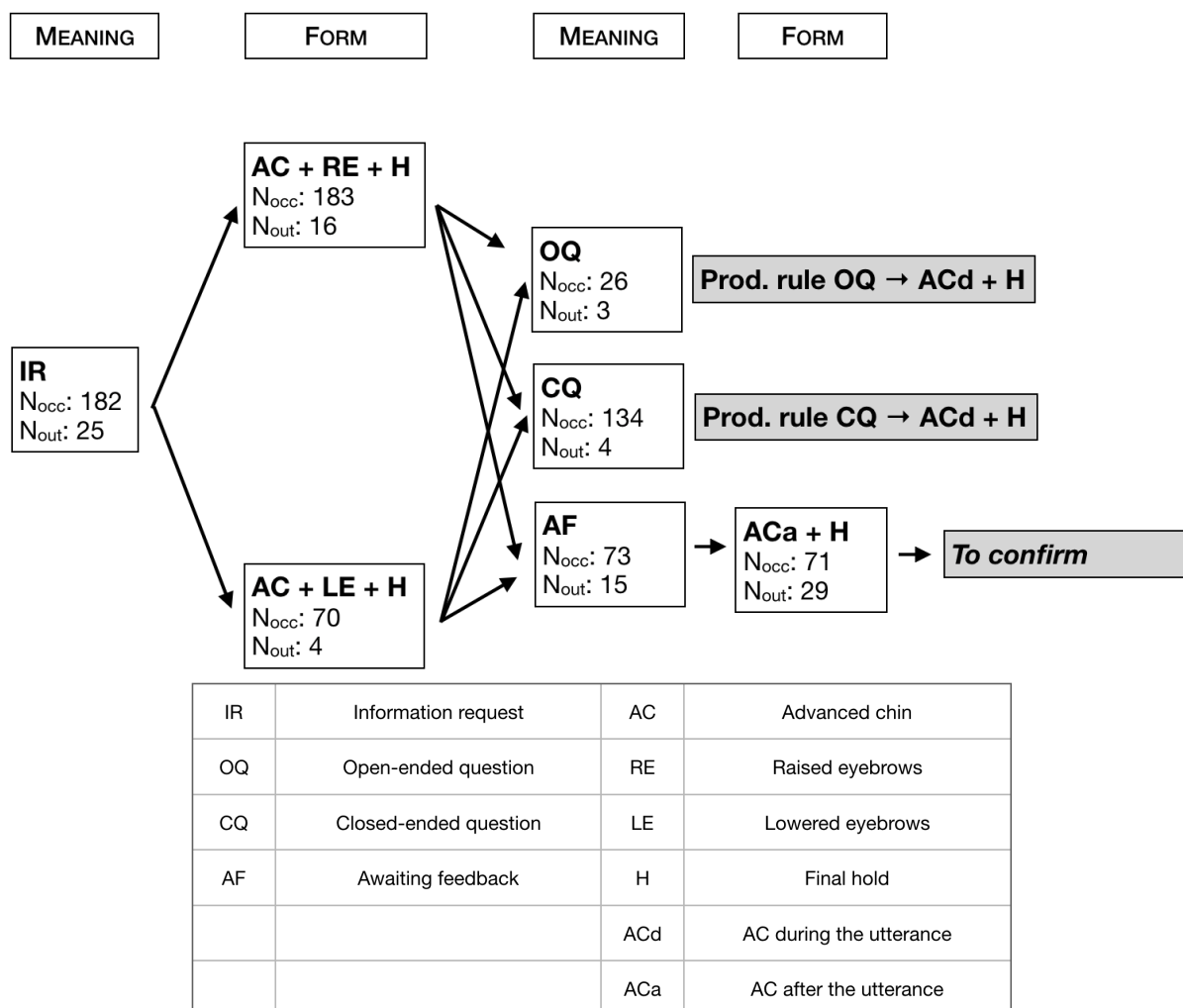


Figure 2: Overview of iterations for IR

## 2. AC + LE + H (61 entries)

- $N_{out} = 25$

Following the methodology, we must now take each of the formed groups separately, because more than one surfaced. For each, we can either recognize a meaning–form association already accounted for by other rules of the known AZee production set, or explore further by going through the steps again, starting with the criterion defining the group. The latter case applies for both groups here, hence two new necessary iterations, one starting with search criterion “AC + RE + H” and the other “AC + LE + H”.

Searching for “AC + RE + H” yields 183 occurrences. To annotate the interpreted meaning, we chose to label open vs. closed questions (“OQ” and “CQ” respectively) as it is reported relevant in the literature.<sup>4</sup>

<sup>4</sup>See footnotes 1 and 2 for a reminder of these two question types.

We also found another type of production, which fell in neither of these two cases: that of the signed flow being suspended by the signer, signifying some form of feedback is awaited (noted “AF”). This contrasts with IR since it is not an *actual* answer that is expected from the addressee. Examples are given below.

### Iteration AC + RE + H (form)

- $N_{occ} = 183$
- Groups found:
  1. **OQ** (15), e.g. “What do you think about going by camper van?”
  2. **CQ** (97), e.g. “Are we going by plane?”
  3. **AF** (55), e.g. “I have two proposals for a trip...”
- $N_{out} = 16$

Since this iteration searched for a form criterion, this could not be a stopping case. We therefore

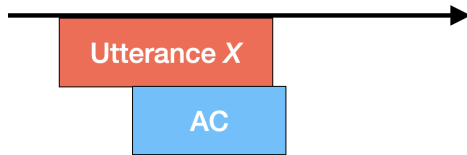


Figure 3: Form synchronization “ACd”: time flows left to right, left boundary of AC between utterance boundaries

continue with iterations until one is reached. The first search is from criterion “OQ”, in which we noticed we could be more specific yet about the timing of the AC form feature. The chin forward movement always started *during* the argument utterance, a refined criterion we note “ACd”, see Figure 3. Together with H, it allowed to capture almost the whole list of occurrences.

#### Iteration OQ (meaning)

- $N_{occ} = 26$
- Single group found: **ACd + H** (23)
- $N_{out} = 3$

This is a possible stopping case, because we find a single form for an identified meaning, and a number of outliers that is low enough ( $< 12\%$ ).

A new production rule named `Open question` can now be defined. It associates meaning OQ with form ACd + H.

Starting with “CQ” yields an outcome similar to the previous with OQ.

#### Iteration CQ (meaning)

- $N_{occ} = 134$
- Single group found: **ACd + H** (130)
- $N_{out} = 4$

This is also a possible stopping case. A new production rule named `Closed question` can be defined. It associates meaning CQ with form ACd + H. We notice that the two new production rules, `Open question` and `Closed question` share an identical form, which is represented on Figure 3. This will be addressed in the results section.

The following iteration, starting with the “AF” criterion, contrasts in form with the prior “OC” and “CQ”. In the case of awaited feedback, the forward chin movement tended to happen *after* the argument utterance. This refined form criterion “ACa” (fig. 4) is now different to ACd (fig. 3).



Figure 4: Form synchronization “ACa”: argument utterance and AC intervals do not overlap

#### Iteration AF (meaning)

- $N_{occ} = 73$
- Single group found: **ACa + H** (58)
- $N_{out} = 15$

$N_{out} = 15$  correspond to occurrences where the chin is not advanced *after* the utterance but rather *during* it. Because it is above the 15% threshold of  $N_{occ}$ , at this point this only shows that 79% of AF occurrences display a chin advancement (AC). We decided to continue the iteration with the highlighted form criterion.

#### Iteration ACa + H (form)

- $N_{occ} = 71$
- Single group found: **AF** (42)
- $N_{out} = 29$

Since this circles back to a previous examined meaning criterion, unsuccessfully searched, and since the present study is focused on IR, we did not pursue this iteration and chose to leave it for future work. This iteration will be done in further studies to confirm or refine with form-meaning association. At this point, results imply that: occurrences ACa are almost always occurrences of AF. However, AF occurrences can have a form that is not ACa.

We now come back to the last unaddressed iteration, namely AC + LE + H, which we processed using the same meaning features as they were equally present.

#### Iteration AC + LE + H (form)

- $N_{occ} = 70$
- Groups found:
  1. **OQ** (11)
  2. **CQ** (37)
  3. **AF** (18)
- $N_{out} = 4$

Each of these three groups are defined by a criterion for which a search has already been performed. So there is nothing here to explore further.



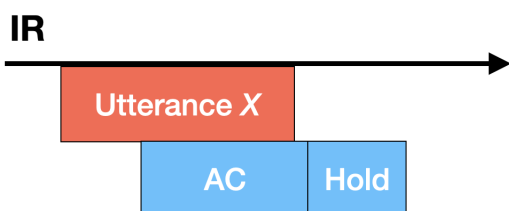


Figure 5: Form synchronization of IR

## 5. Results

Our results point toward two main directions. First, it allows the extraction of two new AZee production rules that we can merge into a single one, as we will explain. Secondly, it emphasizes another phenomenon that will need refinement.

### 5.1. A new production rule

At the end of the iterations, two new production rules have been extracted from the data.

- Open question
  - Meaning: Open question on utterance X
  - Form: Neck starts moving chin forward *during* utterance X (i.e. before the end of X) and final hold
- Closed question
  - Meaning: Close question on utterance X
  - Form: Neck starts moving chin forward *during* utterance X (i.e. before the end of X) and final hold

These two rules share an identical form. This is not a problem, however, a closer look at their respective meanings shows that they could be merged to a more generic production rule. Indeed, from a meaning point of view, “open questions” and “closed questions” are both cases of IR. This leaves us with a single rule, *Information request*, encompassing both and detailed below. It is also illustrated through the synchronization of forms on Figure 5.

- Information request
  - Meaning: Information request through utterance X
  - Form: Neck starts moving chin forward *during* utterance X (i.e. before the end of X) and final hold

### 5.2. Emergence of another phenomenon

The methodology triggered the emergence of a new meaning criterion. It is “Awaiting feedback” (AF), that is actually semantically close to IR.

Indeed, in both cases, the signer is particularly involved in the ongoing interaction. To be more precise, AF semantically differs from IR by the fact that it is not an *actual* answer that is expected by the signer, but rather a simple form of acknowledgment or back-channeling. The form associated to this meaning criterion seems to also be close to IR’s since it involves AC too.

Because that is somewhat beyond the scope of this paper, we decided not to pursue iterations for AF. However, it seems that its form would differ from IR in its synchronization. Indeed, the application of our methodology shows that the unique form associated with IR is “AC *during* the utterance”. Concerning AF, the current state of progress of the methodology does not allow us to establish a new production rule.

## 6. Discussion

This study participated in clarifying the role of eyebrows when requesting an information in LSF. This is an effect of the application of our methodology on SL data.

### 6.1. IR criterion and the role of eyebrows

The methodology applied with starting criterion IR led to exploring two form criteria (“Advanced chin + Raised eyebrows + ‘Final hold’”; “Advanced chin + Lowered eyebrows + ‘Final hold’”) which did not end up to meaning criteria linked with traditional specific question types (i.e. open *versus* close question). Instead, it looped back to IR with a single form cue: “Advanced chin” (AC).

In this respect, these findings do not confirm some claims in the literature on interrogatives concerning an assumed difference of marking between open and closed questions. In fact, no specific shape of eyebrows linked with a particular type of question could be identified. Searching for “closed questions”, we found 72% were produced with raised eyebrows, and 57% for “open questions”. Comparable proportions were also found for lowered eyebrows, regardless of the type of question (42% for open questions; 27% for closed ones).

We extended the analysis to 22 minutes and 32 seconds of supplementary data from Dicta-Sign corpus and found even less relation between the type of question and the form of eyebrows (Table 1). It appears then that the distinction widely reported in the literature is not confirmed by our results.

Type of qu.	RE	LE
CQ.	140 occ. (70%)	58 occ. (29%)
OQ.	21 occ. (61%)	13 occ. (38%)

Table 1: Search for meaning criteria “CQ” and “OQ” extended to a larger data set

## 6.2. Advantages of the methodology

The merging of the two new production rules into a single one shows that the starting criterion, meaning criterion IR, finally made a come back in our search. This is an interesting example of the unpredictable aspect of our method using binary criteria (meaning ones *versus* form ones). It presents the advantage of letting criteria gradually emerge from SL data no matter what the researcher’s intuitions are. Of course, choices are made by the researcher: see for instance the choice to focus on eyebrows in meaning groups AC + RE + H and AC + LE + H, instead of something else. Although this choice was influenced by the literature, we could have chosen to examine something else. Still, it is worth noting that the application of the methodology progressively took us away from this form criterion, revealing it to be irrelevant for meaning criterion IR. At the end, criteria can always be split, or circle back.

This work also led us to define precisely our starting criterion to search in data. If finding occurrences of a given criterion proves to be an ambiguous process, it indicates that the searching criterion is not precise enough and needs refinement. This consideration made us define the meaning criterion "Information request" (IR) instead of the less appropriate "Interrogative" or "Question".

On a larger scale, this allows us to question traditional categories of analysis in linguistics that are frequently elaborated initially for (some) spoken languages. These categories are not necessarily inaccurate but they certainly mold the way we work on SLs. Caution is therefore required when it comes to applying them to the analysis of these languages without prior critical exam. In this, the methodology used in the present work helps bringing out new categories based on the SL data, and undoubtedly more accurate ones to implement for SL generation.

## 7. Conclusion and prospects

Our results do not confirm the hypothesis that eyebrows play a dominant role in requests for information in LSF. We now need to continue applying our method to refine the eyebrows form. Indeed, on 182 occurrences of information requests, 93% ( $n = 170$ ) display a movement of eyebrows (raised or lowered). This is not a problem in our approach

since AZee production rules require only necessary and sufficient elements (see footnote 3).

However, different eyebrow positions might be cues of something different than the type of question. In this regard, following some recent studies, we could focus on the role of potential *biases* in closed requests, such as the signer’s prior belief (Cañas Peña, 2019; Oomen and Roelofsen, 2023). These authors also introduce more fine-grained sub-categories for closed-questions (for instance, *inner* and *outer* closed-questions) that could be tested as meaning criteria in our data. Another hypothesis to explain eyebrow movements is to consider them simply as the result of a stack of other rules, for instance concerning facial expressions and the role of expression of emotions.<sup>5</sup>

We also intend to continue iterations for AF meaning criterion. To do so, it might be useful to extend this analysis to LSF data containing other discourse genre. Indeed, this would allow us to check if the distinction between timing of chin advancement<sup>6</sup> is confirmed by SL data, or if the meaning AF involves another form. Other LSF data could also allow to test close semantic categories such as “Imperative”, which also involves a request of reaction from the interlocutor. In terms of form, we should pay attention as well to the dynamics of the advancement movement in addition to its timing: for instance, a clear-cut one or a progressive one.

Finally, within the frame of AZee, the addition of a new production rule in the AZee production set increases its potential coverage of LSF. The previous set of AZee rules covered 96.1% of LSF discourse in the only corpus entirely represented with AZee, *40-brèves* corpus (Filhol and Challant, 2022), (Challant and Filhol, 2022). As this corpus does not include any questions, we are not yet in a position to quantitatively evaluate the new coverage rate on this corpus. Other studies will follow to further enrich the AZee system, thanks to the ongoing representation with AZee of the *Mocap1* corpora (Benchiheub et al., 2020). This corpus, made up mainly of image descriptions, is a specific register of LSF that represents a challenge in terms of linguistic description and modeling. It is indeed hardly modeled with sole glosses. Our initial findings underline the descriptive potential of the AZee system in this area. This is in line with an increasing coverage of AZee, for LSF for now, but the methodology could also be applied to other SLs data.

<sup>5</sup>For a study on this specific topic in the AZee framework, see Challant and Filhol, *accepted*.

<sup>6</sup>Starting *during* the utterance for IR rather than *after* the utterance for AF.

## 8. Acknowledgement

This work has been funded by the Bpifrance investment “Structuring Projects for Competitiveness” (PSPC), as part of the *Serveur Gestuel project* (IVès and 4Dviews Companies, LISN - Paris-Saclay University, and Gipsa-Lab - Grenoble Alpes University).

The authors also thank Claire Danet for her help in the first steps of this work.

## 9. Bibliographical References

- Charlotte Baker and Dennis Cokely. 1980. *American sign language. A Teacher's Resource Text on Grammar and Culture*. Silver Spring, MD: TJ Publ.
- Sara Cañas Peña. 2019. The marking of polar interrogatives in catalan sign language a first attempt to solve the puzzle. *ConSOLE XXVII*, 1:1.
- Carlo Cecchetto. 2012. *Sentence types*, page 292–315. Mouton De Gruyter.
- Camille Challant and Michael Filhol. 2022. A First Corpus of AZee Discourse Expressions. In *LREC 2022, 13th Conference on Language Resources and Evaluation, Representation and Processing of Sign Languages*, Marseille, France.
- Camille Challant and Michael Filhol. accepted. Extending AZee with Non-manual Gesture Rules for French Sign Language. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Torino, Italy.
- Jane Coerts. 1990. The analysis of interrogatives and negations in sln. In *Proceedings of the Third European Congress on Sign Language Research. Hamburg*, page 265–277.
- Connie de Vos, Els van der Kooij, and Onno Crasborn. 2009. [Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands](#). *Language and Speech*, 52(2–3):315–339.
- Colette Dubuisson, Johanne Boulanger, Jules Desrosiers, and Linda Lelièvre. 1991. [Les mouvements de tête dans les interrogatives en langue des signes québécoise](#). *Revue québécoise de linguistique*, 20(2):93–121.
- Michael Filhol. 2008. [Modèle descriptif des signes pour un traitement automatique des langues des signes](#). Ph.D. thesis, Université Paris Sud-Paris XI.
- Michael Filhol. 2021. [Modélisation, traitement automatique et outillage logiciel des langues des signes](#). Habilitation à diriger des recherches, Université Paris-Saclay.
- Michael Filhol and John McDonald. 2020. The synthesis of complex shape deployments in sign language. In *Proceedings of the 9th workshop on the Representation and Processing of Sign Languages*.
- Michael Filhol and John C. McDonald. 2018. Extending the azee-paula shortcuts to enable natural proform synthesis. In *sign-lang@ LREC 2018*, page 45–52. European Language Resources Association (ELRA).
- Mohamed Hadjadj, Michael Filhol, and Annelies Braffort. 2018. Modeling French Sign Language: a proposal for a semantically compositional system. In *Proceedings of the Language Resources and Evaluation Conference*, page 4253–4258, Miyazaki, Japan.
- Thomas Hanke, Lutz König, Sven Wagner, and Silke Matthes. 2010. Dgs corpus dicta-sign: The hamburg studio setup. In *sign-lang@ LREC 2010*, page 106–109. European Language Resources Association (ELRA).
- Annika Herrmann, Sina Proske, and Elisabeth Volk. 2019. [Question-Answer Pairs in Sign Languages](#), page 96–131. Brill.
- Ulrika Klomp. 2021. [A descriptive grammar of Sign Language of the Netherlands](#). LOT Amsterdam.
- John McDonald, Rosalee Wolfe, Sarah Johnson, Souad Baowidan, Robyn Moncrief, and Ningshan Guo. 2017. [An Improved Framework for Layering Linguistic Processes in Sign Language Generation: Why There Should Never Be a “Brows” Tier](#), volume 10278 of *Lecture Notes in Computer Science*, page 41–54. Springer International Publishing, Cham.
- Bill Moody. 1983. *La langue des signes, Tome 1*, i.v.t. edition. Paris.
- Carol Jan Neidle, Judy Kegl, and Benjamin Bahan. 2000. *The syntax of American Sign Language: Functional categories and hierarchical structure*, cambridge, ma: mit press edition.
- Marloes Oomen and Floris Roelofsen. 2023. [Biased polar questions in sign language of the netherlands. two functions of headshake](#). In *FEAST*, volume 5, page 156–168.
- Karen Petronio and Diane Lillo-Martin. 1997. Wh-movement and the position of spec-cp: Evidence from american sign language. *Language*, 73(1):18–57.

Marie-Anne Sallandre, Anne Zribi-Hertz, and Marie Perini. 2021. *La langue des signes française (Lsf). Projet Langues et Grammaire en Ile-de-France*.

Rachel Sutton-Spence and Bencie Woll. 1999. *The linguistics of British Sign Language: an introduction*. Cambridge University Press.

Traci Patricia Weast. 2008. *Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows*. Ph.D. thesis, University of Texas at Arlington, Arlington.

Traci Patricia Weast. 2011. *American Sign Language Tone and Intonation: A Phonetic Analysis of Eyebrow Properties*, page 203–226. De Gruyter Mouton.

Ulrike Zeshan. 2004. Interrogative constructions in signed languages: Crosslinguistic perspectives. *Language*, 80(1):7–39.

## 10. Language Resource References




Belissen, Valentin and Braffort, Annelies and Gouiffès, Michèle. 2020. *Dicta-Sign-LSF*. v2, ISLRN 442-418-132-318-7. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Benchiheb, Mohamed-El-Fatah and Berret, Bastien and Braffort, Annelies. 2020. *MOCAP1*. v1, ISLRN 502-958-837-267-9. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Filhol, Michael and Challant, Camille. 2022. *40 brèves*. v2, ISLRN 988-557-796-786-3. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.



# Multilingual Synthesis of Depictions through Structured Descriptions of Sign: An Initial Case Study

John McDonald<sup>1,2</sup>, Eleni Efthimiou<sup>2</sup>,  
Stavroula-Evita Fotinea<sup>2</sup>, Rosalee Wolfe<sup>1,2</sup>

<sup>1</sup> School of Computing, DePaul University, Chicago, IL, USA

<sup>2</sup> Institute for Language and Speech Processing, ATHENA Research Center, Athens, Greece  
{jmcdonald, rwolfe}@cs.depaul.edu, {eleni\_e, evita, rosalee.wolfe}@athenarc.gr

## Abstract

Sign language synthesis systems must contend with an enormous variety of possible target languages across the world, and in many locations, such as Europe, the number of sign languages that can be found in a relatively limited geographical area can be surprising. For such a synthesis system to be widely useful, it must not be limited to only one target language. This presents challenges both for the linguistic models and the animation systems that drive these displays. This paper presents a case study for animating discourse in three target languages, French, Greek and German, generated directly from the same base linguistic description. The case study exploits non-lexical constructs in sign, which are more common among sign languages, while providing a first step for synthesizing those aspects that are different. Further, it suggests a possible path forward to exploring whether linguistic structures in one sign language can be exploited in other sign languages, which might be particularly helpful in under-resourced languages.

**Keywords:** Sign Language Synthesis, Avatar, AZee, Geometric Constructions, Multilingual, Translation

## 1. Introduction

Signing avatars have held an, as of yet, unrealized promise both as an assistive technology for bridging Deaf-hearing communication and as an educational tool for Deaf and hearing sign language learners. Even while advances in both computer animation and machine learning are bringing us closer to realizing some of these long-held goals avatars are often eyed with suspicion by the Deaf community. This is due to their failure to legibly portray both the full linguistic structure of signing and the subtleties of human motions (Kipp et al., 2011). Another contributing factor are overly confident claims often made by companies and researchers concerning the capabilities of their avatars (Wolfe et al., 2022) (Deutscher Gehörlosen-Bund et al., 2024).

One significant challenge for wide-scale applicability of signing avatars is the great diversity of sign languages across the world. Ironically, it is a common misconception among the hearing population that sign languages must be “universal”, partially because of the perceived prevalence of iconicity in signing (Hohenberger, 2007). People in the U.S., for example are often surprised to hear that, not only is British Sign Language (BSL) a completely separate sign language from American Sign Language (ASL), but that ASL shares more in common with French Sign Language (LSF) from which it was derived (Fischer, 2015).

More surprising yet is when they hear that Switzerland has three recognized sign languages, Swiss German (DSGS), Swiss French (LSF-SR), and Swiss Italian (LIS-SI), used in different regions and are different both from each other and from

the sign languages of Germany (DGS), France (LSF) and Italy (LIS) (Eberhard et al., 2022). A similar situation can be found across South America where Venezuelan, Honduran and Argentine sign languages are all distinct despite the fact that Spanish is the common spoken language across the region (Akorbi, 2023).

An avatar that signs in only one language will be limited to serving the population of a single region, and several projects have worked towards creating a signing avatar that can communicate in several languages (Efthimiou et al., 2010). However, each of these have been limited in the linguistic features that they can encompass and the naturalness of human motion that they can achieve, both of which decrease the legibility of the resulting synthesized sign. More recently, the EASIER project has taken up this challenge and the present case study has arisen from this work (EASIER-Project, 2024).

Throughout this discussion it will be important to remember that a signing avatar is not the same as a spoken-to-sign translation system. The same is true for a linguistic description system. While it is true that considerations of Deaf-hearing communication cause researchers to focus more often on translations of spoken and written language to sign, most signing happens among native signers in the Deaf community and has no relationship to spoken or written language, nor is there a broadly accepted written form for sign language. It is imperative then to see both the linguistic description and the signing avatar for what they actually are, respectively, a description of the signed discourse, and a display system to communicate that signing visually. Whether the signed discourse arises

from a spoken/written translation or from naturally occurring sign is irrelevant.

This paper will present promising first steps towards animating sign in multiple languages directly from a rich structured description of the desired signing. The descriptions are not tied to translations from spoken languages and can also encompass natural signing that arises between native signers. The hierarchical structure afforded by the linguistic descriptions provides important cues to the animation system that informs nonmanual and prosodic signals which include the relative timing of both manual and non-manual motions (Sandler, 2010). These significantly improve the legibility of the resulting synthesized sign. Furthermore, this effort focuses on forms signing that have traditionally been a challenge for sign synthesis systems, namely classifier constructions and geometric depictions. In fact, it is the very geometric nature of these constructs that makes them more understandable among different sign languages and provides a foundation upon which a multilingual system may be achieved.

Finally, these first steps will point to important ways that linguistic descriptions of sign and avatar animations can interact, allowing each side to learn from the other. Indeed a synthesis system built on hierarchical descriptions of sign and legible animation may provide a powerful tool for studying and testing linguistic theory.

## 2. Lessons from prior multilingual efforts

A brief review of past efforts towards the multilingual display of sign language can highlight many of the challenges faced in such an endeavor, and can also point to strengths in each approach that can be leveraged in the new approach explored here. In the past, the main efforts for sign language representation and display can be divided into three main categories:

1. *Phonetic systems* such as HamNoSys attempt to encode the motions of signing via the fundamental parameters of human posture and motion, such as handshape, palm orientation, movement (Hanke, 2004). For example, large libraries of HamNoSys/SigML annotations were used in the Dictasign project to allow signing in several European sign languages (Dictasign, 2012). Dictionary signs and other gestural units were described to the avatar phonetically in terms of their parametric linguistic labels to enable easy annotation. The supporting avatar was quite flexible in the range of vocabulary it could express, due to the existence of large corpora of annotated signs

in a few languages. However it was limited in the quality of animation output due to the coarseness of the linguistic description, and the lack of prosodic cues included in the animation (Caridakis et al., 2011) (Kipp et al., 2011). Structure beyond the phonetic is necessary for portraying the prosodic structure of the language, which is essential for legibility. Nevertheless, phonetic linguistic notations can provide large repositories of data for an animation system.

2. *Gloss-based systems* rely on a series of glosses, i.e. written words that provide the closest approximation to the meaning of a sign, which dictate the content of the desired signing. A dictionary-based lookup of these concepts from the target sign language is then used. This lookup can be in the form of pre-animated (Wolfe et al., 2011) or prerecorded sequences (Gibet and Marteau, 2023). While very flexible, these efforts suffer from small vocabulary sizes, due to the cost of either animating sign or recording humans with motion capture. In addition, while the signing of individual dictionary entries can be of very high quality in either approach, the process of stitching sequences of these recordings can be stale and awkward if the system has no knowledge about the larger grammatical structures that link them together. This includes both non-manual and timing considerations. Because of this, some more recent gloss based systems have explored adding prosodic and non-manual instructions to gloss streams (Adamo-Villani and Wilbur, 2015) (Hanke et al., 2023); developments which have greatly enhanced the quality of the resulting animation. One of the great lessons from this approach is that the more structure that the representation provides, particularly for prosodic and nonmanual communication, the more legible the synthesized sign will be. In addition, using a library of phonetic description such as HamNoSys to describe each gloss to the avatar could help alleviate the problem of small dictionaries, but at present efforts to animate directly from such sparse linguistic data remains problematic due to the robotic nature of the resulting motion.
3. *Deep-learning systems* which exploit large libraries of annotated video and/or motion capture recordings of sign, and attempt to produce video or skeletal motion directly from the desired spoken text (Saunders et al., 2021). These efforts have explored multilingual display in British Sign Language (BSL), the Sign Language of the Netherlands (NGT) and DGS. The major current challenge for these tech-

niques is the size of available corpora. Among the largest annotated corpus between signed and spoken languages is the DGS Corpus (Hanke et al., 2020), which contains in excess of 63,000 pairs. This may seem large, but pales in comparison to the roughly 15 billion pairs that are exploited for modern translation systems between spoken languages. These efforts also suffer from a major issue when it comes to linguistic study. Since the neural networks that drive these systems are largely black-boxes that produce an animation with little indication of how the system is producing the result, it can be difficult to derive meaningful linguistic data on the structure of the resulting signed discourse. Nevertheless, there is no denying the power of deep learning techniques, and as both the corpora and the techniques that exploit them advance, they will no doubt provide increasingly important for informing animation tasks (Choudhury, 2022).

This last issue of corpus size can be particularly problematic for so-called under-resourced sign languages. Compared to spoken languages, most NLP and Deep Learning efforts would consider all sign languages under resourced (Börstell, 2023), however, even among sign languages there is a great disparity between the resources amassed for languages like ASL, DGS, and BSL, and those that have been gathered for languages with smaller communities such as Greek Sign Language (GSL), for which corpus sizes and native populations are significantly smaller. Any effort that aims to deal with multilingual display must have a method for handling such disparities.

### 3. Challenges and opportunities

The misconception that sign language should be universal does arise out of two interesting aspects of both sign languages and their native users. First, there is evidence that signers, as opposed to users of spoken language, are often more adept at interlingual communication. One factor may be due to the continual practice signers get when attempting to communicate with both hearing people and Deaf people from other cultures (Sacks, 2022).

Another important factor is the very nature of the languages that they are fluent in. There tend to be more aspects of signing that are found to be shared between sign languages than is the case for spoken languages. One of these is the form in which an utterance can mimic the shape, motion or sound that is being conveyed (Perlman et al., 2018).<sup>1</sup> The visual structure of sign includes geometric

<sup>1</sup>In spoken language onomatopoeia is an example of iconicity, and some sources (Perlman et al., 2018)

constructs such as classifier predicates, size and shape specifiers and depicting signs that use the body in geometric ways (Zwitserslood, 2012).

In fact, it is partly due to the prevalence of certain types of geometric constructs, which are similar across sign languages, that has led to the greater success of International Sign compared to similar efforts in spoken language (Mesch, 2010). In particular, International Sign often uses classifiers and depictions as a more interlingually understandable way to communicate objects and actions than fixed signs (McKee and Napier, 2002). The present effort will seek to exploit these aspects in an effort to build a first step towards a multilingual display.

#### 3.1. Classifiers and depictions

Geometric constructions, including size and shape specifiers and classifier predicates that depict the placement and movements of objects are observed in most sign languages, and, while the specific handshapes differ significantly between sign languages, the motions of the body that depict the placement and movement are largely similar (Pfau et al., 2012). For example, when placing a small round object like a plate, many languages use the hands to mimic the shape of the object, and then use a downward motion to place that object figuratively in space relative to other objects that may be depicted. Unless the object is hanging on a wall or on the ceiling, this motion will naturally be downward. More generally, the placement of the object will be expressed naturally by a movement toward the surface that the object is resting on.

Extensive examples of geometric constructs like these may be found in the Mocap1 corpus in LSF from the LISN (formerly LIMSI) laboratory (LIMSI et al., 2022) (Benchiheub et al., 2016). In this corpus, Deaf participants were provided with pictorial stimuli that they were then free to describe in any way they wish. For example, one stimulus was the picture of the neatly-decorated dining room shown in figure 1. Descriptions of the room varied significantly between participants with some describing the room very sparsely and others in great detail.

One key characteristic of classifier constructions in sign language is that they are among the least “lexical” parts of signing that occur in native discourse in the sense that signers will often use very few dictionary signs when describing either the action or structure of a scene. In a 30-second section of one Mocap1 participant’s description, the signer described the table setting in great detail. The entire sequence, however involves only seven dictionary signs, which in English would be glossed RUG, TABLE, CHAIR, PLATE, GLASS, KNIFE and

(Handspeak, 2) have applied versions of the term onomatopoeia to signing as well in place of iconicity





Figure 1: Mocap1 stimulus for the description of a dining room

FORK. The rest of the signing is dedicated to the geometric placement of the rug on the floor, the table on top of the rug, two pairs of chairs facing each other at the table, four plates arranged symmetrically on top of the table with the four glasses and the four pairs of knives and forks arranged around the plates. Figure 2 contains examples of the signer placing a plate, two glasses and a knife and fork pair.

It is precisely due to the fact that classifier predicates show a geometric placement or movement of an object that, once the object's type is established, the action mimics the natural ways in which the object settles onto a surface or moves in space. This makes it far more likely that these motions will be similar among sign languages. Note, however, that it is not claimed here that they are precisely the same in all sign languages, but that their similarity gives a starting point to work from for multilingual display. The system described below will have the flexibility to accommodate such differences. For example, the classifier for a moving vehicle like a car is signified in ASL by an "Three" handshape oriented on its side with the ulnar side of the palm against the surface, whereas in LSF, it is indicated by a flat hand with the palm flat on the surface. In both instances, the extended fingers indicate the direction that the car is facing, though this may not always be the case for other language pairs<sup>2</sup>.

### 3.2. Sign language description

Every synthesis system must have, at its core, some method of describing the desired signing, and we must consider the lessons from prior efforts discussed in section 2, when choosing the description system. Of primary consideration for our case study here is the legibility of the resulting synthesis, and as has been seen in prior work, non-manual signals that give purposeful motion to

<sup>2</sup>There is also usually a difference between the classifier for a chair in LSF and ASL, but the signer in this LSF example actually uses a handshape very close to the ASL version, which is usually used in LSF for a small animal

the spine, head and face are key to the quality of the resulting signing. Further, the description system must be capable of informing varied timing and pacing for movements that are key to breaking up the robotic monotony that have plagued past efforts to synthesize sign from phonetic descriptions. In this respect, the AZee description system (Filhol et al., 2014) has proven to be a powerful tool for describing, not only the basic gestural units of signing, but also the connecting structure that provides necessary nonmanual and timing information (Filhol et al., 2017).

Another key aspect of AZee from a synthesis perspective is that it has proven extremely capable in its ability to describe classifier constructions and depictions for avatar synthesis, where this table description was animated directly from the AZee description in LSF (McDonald and Filhol, 2021). The only elements that were supplied by an artist were the seven animations of the citation forms of RUG, TABLE, etc, and single example poses for each of the classifiers. We will not review the AZee description in its entirety, but will recall the main AZee rules that are used in the description, and select examples of how those rules are used. The names and parameters of the rules have been updated in accordance with the latest published AZee notation (Filhol et al., 2024).

- *in-context(context, process)*, formerly *context*, this rule provides the main glue that knits sections of signing together. It causes a hold to happen at the end of the *context* along with a blink. It indicates that the signing described by *process* is to be understood in the context of the signing *context* that comes before it. For example,

*in-context(table placed on rug, items on table)*  
(1)

indicates that the table placed in the scene is where the list of items is placed. This is often one of the top-level rules that builds the hierarchy of signed discourse, and both *context* and *process* are often large descriptions of signing themselves.

- *instance-of(type, element)*, formerly *category*, this rule indicates that the signing described by *element* is to be understood as an instance of *type*, which comes before it. The signing in *type* is accompanied by a subtle raising of the eyebrows, a tilting of the head and a short transition between the two with no hold on the *type*. For example,

*instance-of(glass, placements of cylinders)*  
(2)

indicates that the cylindrical objects being placed on the table are glasses.





Figure 2: Placements in the Table Description

- *place-object*(*loc, class*), is a rule that indicates the placement of an object in signing space. Here *class* is a classifier that indicates what kind of object is being placed at *loc*, which does not have to be one hand. The signer forms the classifier and with a downward settling motion with normal speed, places it at the *loc*. An example of this is

*place-object*(*Midssp, prf-cylindrical-small*) (3)

in which the signer’s hand assumes C-handshape oriented vertically, figuratively around the cylinder, held just above the point, which here is “middle of signing space”, and then settles the hand down to that point.

- *each-of*(*items*), here *items* is a list of things to be signed with emphasis on each individual element of the list. Each element is signed at a normal speed, the last posture of each element is held and a medium transition time is used between the items. An example of this is

*each-of*(*two pairs of glasses*) (4)

The signer places a pair of glasses simultaneously, one with each hand, then pauses before placing another pair.

- *all-of*(*items*), again *items* is a list of things to be signed, but in contrast with *each-of*, focus is placed on the group of elements from the list. Each element is signed at a faster than normal and the motion bounces at the end of each element instead of holding. An example of this is

*all-of*(*four flat round plates*) (5)

In this case, the signer uses both hands to show the round shape of the plate and a downward motion to place it, but instead of pausing, then proceeds to rapidly move with a bouncing motion to place the remaining three. The effect is seen as a group of plates as noted in (McDonald and Filhol, 2021).

One important aspect of AZee is that it does not presume a definition of “lexical” or “fixed” signs, even though classical linguistic models of signing would clearly indicate them in this discourse. For example, consider the signing GLASS before the classifier placements. In LSF, GLASS corresponds to an AZee rule defining that a C-handshape on the dominant hand is tapped twice vertically on the flat palm of the non-dominant hand. Since this is linked with an *instance-of* rule, the sign GLASS is performed while raising the eyebrows and chin, followed by the classifier placements, which situate the glasses in space. This same kind of pairing occurs in other sign languages as well. For example, in videos explaining ASL classifiers (ASL-That, 2012) (Handspeak, 1), the demonstrators display remarkably similar eyebrow and/or head motions after each lexical sign and before each classifier placement in mid signing space.

In this scene and in others that have been animated with the signing avatar Paula (Wolfe et al., 2011), the lowest common denominator among rules in AZee that seem to correspond to what linguistics would normally call a “lexical” sign is that it can be applied in the AZee description without any parameters at all, i.e. it can be signed generically. This cannot be said of any of the other rules listed above, all of which require parameters to be specified. For example, *place-object* must have both the classifier and location specified. There is no reasonable choice for a “default” object or “location” from AZee’s perspective. This is different from the signs for RUG, TABLE, PLATE, etc., which can all be signed in a generic citation form. We will use the term “lexical” for AZee in this sense throughout this paper.

#### 4. Exploiting common structures

The goal of synthesizing the same discourse in several languages can, of course, be seen as a translation task between sign languages, as in the present case study since the AZee description for the signing in question came from LSF. Our goal

is then to take the description of signing in LSF that describes a setting of a table, and produce an equivalent understandable discourse in other languages. For this case study, we chose to work with DGS and GSL as the other target languages. The key is to look for structures in signing that they would seem to have in common.

Discussions with native signers, analysis of examples from corpora in both DGS and GSL, and the theoretical reasoning above concerning the strong similarities of placements and movements, reveal that classifier placements are likely to be done in a very similar manner across these sign languages. For the relative placement of object in a scene, while the specific classifier handshapes may differ in form, all three of these languages place objects using a downward settling motion. Since a large majority of the signing in discourse such as this description fall into this category, the current discourse is a good candidate for this case study, and we will begin with the following hypothesis:

**Hypothesis 1** *AZee rules such as place-object and move-object have a similar form in each of these target languages.*

This hypothesis does, as we have noted, seem to have support in current linguistic theory. Note, however, that the geometrical presentation of its parameter *class* will generally not be the same, but we will assume that we can find a classifier for each object in question.

On the other end of the scale, lexical signs are often very different across sign languages, as can be seen with a casual inspection of sign dictionaries in any two sign languages. Even languages like LSF and ASL, which do share a common history and do have some cognates, are generally so different in their lexicons as to render many signs inscrutable to signers of the other language. From this, we can conclude that some mechanism will need to be used to animate the concept that corresponds to a lexical item in each sign language. Note that we will not assume here that it is possible in all cases to translate lexical signs in one language to single *lexical* signs in another, but the system will have to be flexible enough to handle this transfer.

Somewhere in the middle, we have all the processes that link these signs together, which we will call here *connecting rules*. In LSF, we have *instance-of*, which links each object's lexical sign to its corresponding placements, while whole sections of the discourse, such as the placement of the table as being the area where the rest of the setting is placed, is accomplished via *in-context* rules. Note that these were not arbitrarily chosen but rather the visual cues for these rules, such as eyebrow raises, head-tilts, pauses and blinks, were observed in the signer's movement. Here, there is a huge question concerning whether these kinds of linkages will be

present in other languages.

As mentioned, something similar to the *instance-of* rule has been observed in ASL, and there are instances in other languages. For example, consider this example from the DGS Corpus, <http://tinyurl.com/5chseh3u>, where signer is signing "Das ganze Land, Deutschland", which begins with a shape specifier and the sign for LAND, accompanied by a similar raising of the head and eyebrows. The eyebrows are then lowered for the signing of DEUTSCH. This is similar to the motion described by *instance-of* in LSF. At least in this and several other instances, contextualization seems to be communicated in a similar way to LSF. Thus we will formulate a second hypothesis upon which this case study is based.

**Hypothesis 2** *The connecting rules instance-of, in-context, each-of, all-of have a similar form in each of the target languages.*

Again, we are not claiming here that this AZee rule applies in general in DGS, but that it is perhaps reasonable to try synthesizing DGS discourse with this same AZee rules. Certainly, more study will be needed to confirm or disprove this. The trouble is that it has taken nearly ten years of corpus study to arrive at the current list of AZee rules for LSF. So, the question arises of whether there could be any transfer of learning that can shorten the time required to formulate AZee rules in other languages. This brings up an interesting possibility, which we formalize as a hypothesis that should be tested in the future.

**Hypothesis 3** *If it is possible to synthesize sufficiently high quality animation using these descriptions, then the resulting videos could be tested with the Deaf community for their legibility and fidelity of the message. Furthermore, since the avatar system is generative and able to include or exclude features as necessary, individual linguistic features could be tested.*

When evaluating this hypothesis, it is important to note that such testing would be very difficult with motion capture or video of live signers because it is impossible for a human to reproduce a production precisely while including, omitting or changing only one linguistic parameter. Note also that the application of this last hypothesis is one way to test hypotheses 1 and 2

## 5. Avatar support

The goal of this case study is to accomplish the display of this table-description discourse in multiple languages, based on the same description in AZee. Following the discussion of the last section concerning AZee representations, and in particular Hypothesis 2, the current case study will use

the connecting AZee rules *instance-of*, *in-context*, *each-of*, *all-of* and *place-object* as provided in the current AZee specification. Further study will show if changes to these rules will in general be necessary for a particular target language, or whether they need to be replaced by other connecting rules in a particular target.

Avatar synthesis of the signing will be accomplished using the previously published AZee-Paula bridge (McDonald and Filhol, 2021). This prior work also details how this table scene was synthesized in LSF. To extend this bridge to support animation in the target languages, there are two main classes of rules that need to be transformed into the target languages, GSL and DGS:

1. The seven lexical items, *rug*, *table*, *chair*, *plate*, *glass*, *knife*, *fork* which in this scene are signed essentially in their citation form.
2. Classifier specifications for *class-generic-object*<sup>3</sup>, *class-flat-round-large*, *class-cylindrical-small*, *class-straight-elongated* and *class-large-rectangle*.

For the lexical items, recall that the AZee-Paula bridge contains a system of shortcuts which map such rules to pre-animated sequences on Paula which can then be blended with the motions defined by the connecting rules. In the target languages chosen, fixed signs exist to encompass these concepts and so the transformation is as simple as mapping these rules to the corresponding glosses in the target language. It is important to note that this may not always be the case, and a more complicated transfer may be necessary. This does not occur for the chosen target languages in this table description, and extensions to this mechanism will be explored in the future.

For the classifiers, recall that the bridge uses a system of artist exemplars that set up the configuration of the hand and arm which provides an example posture to build the classifier placement or motion from. The system then uses inverse kinematics to accomplish the placement or movement while using the data from the artist template to inform choices for redundant degrees of freedom (McDonald and Filhol, 2021). Again, all that is necessary is that Paula's database in the target language provides an artist shortcut for each of the classifiers used in the discourse. These may be very different from the ones provided for LSF, and may eventually involve changes to the AZee rules themselves in the target language. However, this does not occur in this case study's table scene description.

---

<sup>3</sup>The signer in LSF used a non-standard classifier for chair which is usually used instead for generic objects, but which is incidentally very close to the classifier used in ASL, DGS and GSL for a chair or a sitting person

## 6. Results

The main database for a language in the Paula system contains all of the information that the avatar needs to animate sign directly from the linguistic description, including pre-animated citation forms of lexical signs and classifier definitions. The main task on the avatar side of this case study was two-fold

- Animate each of the seven lexical items in DGS and GSL in their corresponding databases.
- Set the AZee shortcuts for the rules corresponding to these lexical items to link to these new animations in each database.

Both of these steps were completed for all three languages so that the correct animation in each language was automatically triggered as a shortcut to the corresponding AZee rule. Figures 3 and 4 contain still frames from two of the seven lexical items in the three languages.

The other main difference that can certainly exist between sign languages is the form of classifiers, and thus it is in general necessary to create artist exemplars of each classifier. As part of this initial study, we showed the set of classifier forms to a small group of sign language experts from both the University of Hamburg and the Athena Research Center in Athens. Both groups indicated that the classifiers that were used in the LSF discourse were acceptable for their respective sign languages. Thus the original classifier exemplars built for LSF were transferred to the GSL and DGS databases to perform the synthesis. With these changes, the Paula system could synthesize the table scene description directly and automatically from the original AZee description. The results are displayed in the accompanying video which can be accessed at <http://tinyurl.com/2ch2bwvm>.

## 7. Conclusion and future work

This paper presents a first step towards a multilingual display system in the form of a case study, and as such is limited in scope, but nevertheless points the way towards further development. As noted above, there are several avenues that must be investigated:

1. For this line of research to proceed, the resulting synthesized sign must be tested with the Deaf community. Testing with fluent signers is essential for both formative and evaluative feedback on the system, and will be a key element of testing the three hypotheses outlined in section 4.



Figure 3: Images from the animations for RUG in the three languages

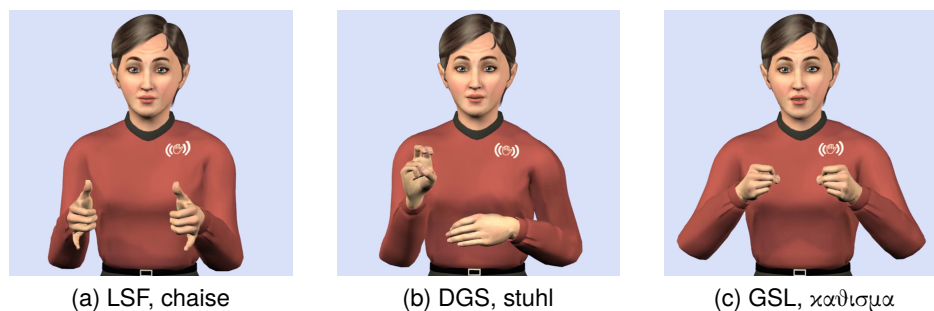



Figure 4: Images from the animations for CHAIR in the three languages

2. One of the main roadblocks is vocabulary acquisition in each of the target languages, however, if signing of sufficiently high quality could be derived from HamNoSys or another phonetic description, existing corpora could address this lack.
3. Classifiers may differ significantly, not only in handshape, but also in the orientation of the hand during placement and movement. An example of this is the classifiers for vehicles in ASL and LSF. Paula can handle part of this difference with the artist exemplar, but changes to the AZee rule for the classifier will also be necessary.
4. The forms of the linking rules described here may differ in other languages, and completely different linking rules may be discovered that don't translate between languages.

Pursuing each of these, with repeated user testing with the Deaf community will point the way to extending this effort to more general multilingual display of sign language.

### Acknowledgement

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982. 

## 8. Bibliographical References

### References

- Nicoletta Adamo-Villani and Ronnie B Wilbur. 2015. ASL-pro: American sign language animation with prosodic elements. In *Universal Access in Human-Computer Interaction. Access to Interaction: 9th International Conference, UAHCI 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part II 9*, pages 307–318. Springer.
- Akorbi. 2023. Is sign language universal? <https://akorbi.com/blog/is-sign-language-universal/>, Accessed: 2024-02-05.
- ASL-That. 2012. ASL Classifiers (CLs) for Furniture & Objects. <https://www.youtube.com/watch?v=xPb8AD1rON0>, Accessed: 2024-02-11.
- Mohamed-El-Fatah Benchiheub, Bastien Berret, and Annelies Braffort. 2016. Collecting and analysing a motion-capture corpus of french sign language. In *sign-lang@ LREC 2016*, pages 7–12. European Language Resources Association (ELRA).
- Carl Börstell. 2023. Ableist Language Teaching over Sign Language Research. In *Proceedings of*



- the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023), pages 1–10.
- George Caridakis, Stylianos Asteriadis, and Kostas Karpouzis. 2011. Non-manual cues in automatic sign language recognition. In *Proceedings of the 4th international conference on pervasive technologies related to assistive environments*, pages 1–4.
- Shatabdi Choudhury. 2022. Analysis of Torso Movement for Signing Avatar Using Deep Learning. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 7–12.
- Deutscher Gehörlosen-Bund et. al. 2024. "sofortiger handlungsbedarf" – deutsche gehörlosendachverbände gegen avatare. <https://www.taubenschlag.de/2024/02/sofortiger-handlungsbedarf-deutsche-gehhoerlosen-dachverbaende-gegen-avatare/>, Accessed: 2024-02-23.
- Dictasign. 2012. Sign language recognition, generation and modelling with application in deaf communication. <https://cordis.europa.eu/project/id/231135>, Accessed: 2024-02-11.
- EASIER-Project. 2024. Intelligent automatic sign language translation. DOI:<https://doi.org/10.3030/101016982>, Accessed: 2024-02-11.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2022. *Ethnologue: Languages of the World*, 25 edition. SIL International, Dallas.
- Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Goudenove. 2010. DICTA-SIGN: sign language recognition, generation and modelling with application in deaf communication. In *sign-lang@LREC 2010*, pages 80–83. European Language Resources Association (ELRA).
- Michael Filhol, Mohamed Nassime Hadjadj, and Annick Choisier. 2014. Non-manual features: the right to indifference. In *sign-lang@LREC 2014*, pages 49–54. European Language Resources Association (ELRA).
- Michael Filhol, John McDonald, and Rosalee Wolfe. 2017. Synthesizing sign language by connecting linguistically structured descriptions to a multi-track animation system. In *International Conference on Universal Access in Human-Computer Interaction*, pages 27–40. Springer.
- Michael Filhol, Thomas von Aschenberg, and John McDonald. 2024. Final Integration Allowing Post-Editing with Discourse Representation Diagrams. <https://www.project-easier.eu/deliverables/>.
- Susan D Fischer. 2015. Sign languages in their historical context. In *The Routledge handbook of historical linguistics*, pages 442–465. Routledge.
- Sylvie Gibet and Pierre-François Marteau. 2023. A Text-To-SL Synthesis System Using 3D Avatar Technology. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.
- Handspeak. 1. ASL Classifiers: [objects] on a table. <https://www.youtube.com/watch?v=GF0iqbD64TA>, Accessed: 2024-02-11.
- Handspeak. 2. "onomatopoeia" in sign language. <https://www.handspeak.com/word/4056/>, Accessed: 2024-02-11.
- Thomas Hanke. 2004. HamNoSys-representing sign language data in language resources and language processing contexts. In *LREC*, volume 4, pages 1–6.
- Thomas Hanke, Lutz König, Reiner Konrad, Maria Kopf, Marc Schulder, and Rosalee Wolfe. 2023. EASIER Notation—a proposal for a gloss-based scripting language for sign language generation based on lexical data. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.
- Thomas Hanke, Marc Schulder, Reiner Konrad, and Elena Jahn. 2020. Extending the Public DGS Corpus in size and depth. In *sign-lang@LREC 2020*, pages 75–82. European Language Resources Association (ELRA).
- Annette Hohenberger. 2007. The possible range of variation between sign languages: Universal grammar, modality, and typological aspects. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 188:341.
- Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*, pages 113–126. Springer.
- LIMSI, CIAMS, and LISN. 2022. **MOCAP1**. OR-TOLANG (Open Resources and TOols for LANGuage) –[www.ortolang.fr](http://www.ortolang.fr).

- John McDonald and Michael Filhol. 2021. Natural synthesis of productive forms from structured descriptions of sign language. *Machine Translation*, 35(3):363–386.
- Rachel Locker McKee and Jemina Napier. 2002. Interpreting into international sign pidgin: An analysis. *Sign language & linguistics*, 5(1):27–54.
- Johanna Mesch. 2010. Perspectives on the concept and definition of International Sign. *World Federation of the Deaf*.
- Marcus Perlman, Hannah Little, Bill Thompson, and Robin L Thompson. 2018. Iconicity in signed and spoken vocabulary: a comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in psychology*, 9:1433.
- Roland Pfau, Markus Steinbach, and Bencie Woll. 2012. *Sign language: An international handbook*, volume 37. Walter de Gruyter.
- Oliver Sacks. 2022. *Seeing voices: A journey into the world of the deaf*. Neha Publishers & Distributors.
- Wendy Sandler. 2010. Prosody and syntax in sign languages. *Transactions of the philological society*, 108(3):298–328.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135.
- Rosalee Wolfe, John McDonald, and Jerry Schnepp. 2011. [An Avatar to Depict Sign Language: Building from Reusable Hand Animation](#). In *International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Berlin, Germany.
- Rosalee Wolfe, John C McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: A question of representation. *Information*, 13(4):206.
- Inge Zwitterlood. 2012. Classifiers. In *Sign language: An international handbook*. De Gruyter.

# Swedish Sign Language Resources from a User's Perspective

Johanna Mesch<sup>1</sup>, Thomas Björkstrand<sup>1</sup>, Eira Balkstam<sup>1</sup>  
Patrick Hansson, Nikolaus Riemer Kankkonen<sup>1</sup>

Stockholm University

{johanna.mesch, bjorkstrand, eira.balkstam, patrick.hansson, nikolaus.kankkonen}@ling.su.se

## Abstract

The Swedish Sign Language Dictionary [Svenskt teckenspråkslexikon] is one of the most visited websites at Stockholm University, with four million visits each year. The dictionary is an easy-to-use resource for the community, families, relatives, students, educators, researchers and other stakeholders that can be accessed through the website, app, and mobile platforms. STS-korpus is an online interface for the Swedish Sign Language Corpus that is linked to the STS Dictionary, enhancing its utility. Other applications, like TSP Quiz and the STS transcription tool, will also be evaluated. In January 2024, we conducted a survey to explore how users utilise Swedish sign language resources in their everyday lives, studies and work, regardless of hearing status and sign language skills. The purpose is to evaluate these resources from a user's perspective, including aspects such as user-friendliness, relevance, comprehensibility and effectiveness in aiding language learning or communication.

**Keywords:** language resource, evaluation, Swedish Sign Language

## 1. Introduction

One crucial aspect of any language resource is how representative it is of the language or languages it covers. Because of this, the resources also serve as language documentation. Sign language dictionaries are essential language resources to meet the needs of sign language interpreters, teachers, students, deaf communities, individuals with special needs, researchers and other users of sign language, not least to elevate the status of sign language (McKee and Vale, 2023). Signed languages are both visual and tactile and they play a significant role in certain communities. Developing sign language dictionaries has been an essential theme, and there are many different methodologies (e.g. (Bragg et al., 2015; Vlášková and Strachoňová, 2021; Schembri and Cormier, 2022; Mesch et al., 2012a). In this paper, we aim to improve our understanding of language resources for Swedish Sign Language (STS) by collecting feedback from users within sign language communities in Sweden. Specifically, we focus on enhancing the utility of the STS Dictionary and STS Corpus. By engaging with users, we seek to ensure that these resources effectively serve the needs of the community and contribute to the advancement of sign language communication.

### 1.1. Swedish Sign Language Dictionary

The STS Dictionary has served as the primary lexical database for Swedish Sign Language since its online launch in 2008. It is based on Brita Bergman's earlier phonological description and transcription efforts (Bergman, 1979). The STS Dictionary, *Svenskt teckenspråkslexikon*

(2024), has been under development for an extended period and now serves as an online video dictionary, currently containing 21,000 entries and 6,700 sentence examples (Mesch et al., 2023). The latest version with several search functions was launched in May 2023 (see Figure 1). Each dictionary entry is represented by a video of the sign or phrase, sentence examples, a Swedish translation, phonological information, phonological variants and internal cross-links to phonologically or semantically equivalent signs – i.e., homophones and synonyms. The etymological description is also added. The STS Dictionary is enhanced by a quiz designed for studying sign language. A transcription tool is also available <https://teckensprakslexikon.su.se/information/om-lexikonet>. New signs are added to the lexical database on an ongoing basis. Crowdsourcing is crucial to developing and improving the STS Dictionary in collaboration with the deaf and sign language community (Riemer Kankkonen et al., 2018). The STS Dictionary also contributes to the Multilingual Sign Language Wordnet but this is not dealt with in the survey because it is still in the early stages of development (Bigard et al., 2024). Complementing the STS Dictionary is the STS Corpus, which emerged from two projects conducted between 2003 and 2004 and 2009 and 2011 (Mesch, 2023). These language resources are freely available and interconnected, facilitating access and usability for users within the Swedish Sign Language community who are seeking comprehensive information and linguistic analysis.

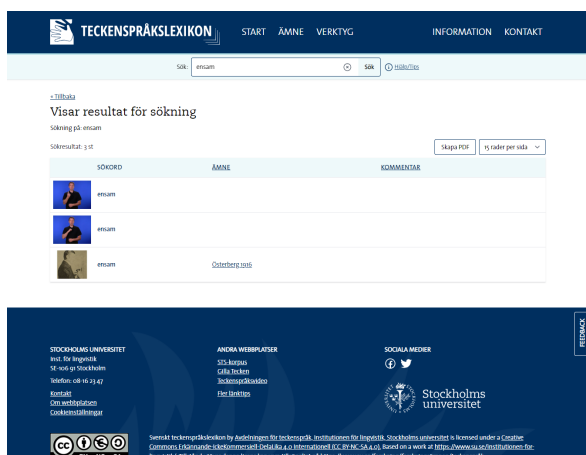


Figure 1: Swedish Sign Language Dictionary

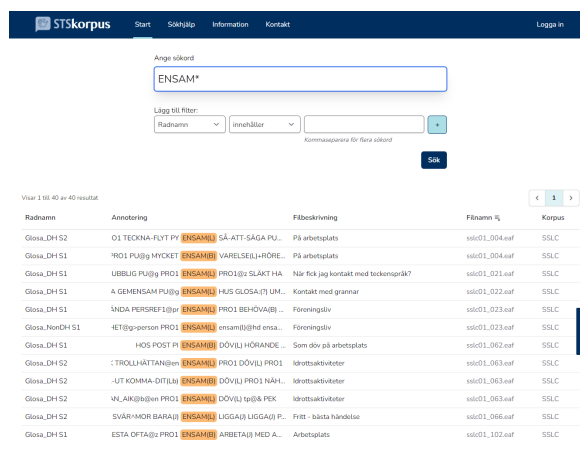


Figure 2: STS-korpus

## 1.2. STS Corpus

The Swedish Sign Language Corpus has been available via the online tool STS-korpus, [teckenspråkskorpus](https://teckenspråkskorpus.su.se) (2024), since 2020 (Figure 2). The STS Corpus consists of 24 hours of video data (conversations, narratives, and presentations) from 42 different signers from three regions of Sweden collected during the period 2009–2011 (Mesch et al., 2012b). All 24 hours of data have been annotated with sign glosses and thus far approximately 76 percent of data has been translated into Swedish (free translation). The annotation work continues, and annotation conversations are still developing. A limited part of the collection of 190,000 sign tokens from this corpus is accessible online via STS-korpus at <https://teckenspråkskorpus.su.se> (Öqvist et al., 2020). In addition to the existing data, the STS Corpus has integrated additional contributions from SSLC2017, including data from six young signers. Moreover, the corpus now includes the updated version of data from the European Cultural Heritage Online (ECHO) corpus for STS from 2004 and the Tactile Swedish Sign Language Corpus from 2016. As part of ongoing efforts, old video recordings and pedagogical materials are continuously being incorporated into the corpus, further enhancing its depth and breadth (Mesch, 2023). Previously developed and managed independently, the STS Dictionary and STS Corpus are now interlinked. It is this collaboration between those working on the STS Dictionary and the STS Corpus that led to the creation of the web-based tool STS-korpus, which facilitates seamless integration between these two resources (Öqvist et al., 2020).

## 1.3. STS Resources as Digital Tools in Educational Settings

Digital methods play a crucial role in sign language education and interpreting across all levels, including the first, second and third cycles of higher education. In 2013, Stockholm University launched the first three-year bachelor's programme in sign language and interpretation, which offers a blend of theoretical and practical courses focused on STS and interpretation between Swedish and STS. The programme emphasises practical learning, aiming to help students enhance their language skills from beginner to advanced level in both STS production and comprehension. Students begin using the STS Dictionary as a digital tool from the first semester of their first year. By the third semester of their second year, they are incorporating the STS Corpus into their studies (Björkstrand et al., 2022). These language resources are invaluable tools for students throughout their academic journey, allowing them to develop their skills progressively and at their own pace through study and practice (Leeson et al., 2019).

## 1.4. Linking of Two STS Language Resources

Here, we briefly present the two language resources and their linking to each other to enhance the functionality of the resources (Öqvist et al., 2020). Collaboration between the dictionary and the corpus pertains to lexical issues, primarily sign lemmatisation. Another significant issue concerns depicting (classifier, non-lexical signs) signs, which are difficult to describe in the STS Dictionary because it requires establishing limited meanings for the signs that do not align with their usage, as discussed in several earlier publications, e.g., (Johnston, 2008). This sign category exists in the corpus but not yet in the dictionary, and collaboration be-



tween the two resources can be used to address this issue. Only some partially lexicalized signs have been added to the STS Dictionary.

The STS Corpus and STS Dictionary are connected through reciprocal linking. The links within the corpus show users summarised information from the dictionary, including a sign video, a description of the sign and a transcription of the sign. There is also an option to go directly to the associated entry in the dictionary in order to learn more. Similarly, dictionary entries for signs that appear in the corpus show the total number of times the sign is used in the corpus. When clicked on, this information takes the user to a list within the corpus that includes every occurrence, to explore further. The integration also has a more systematic feature called “missing corpus”. Each night, an automated process cross-references every annotated term in the corpus with the dictionary database. Before this comparison, terms are checked against a global ignore list, which excludes unhelpful terms from the analysis. This process helps us find discrepancies and refine and enhance the data in both systems, such as missing glosses in the dictionary and misannotated terms within the corpus. As the discrepancies are rectified, the dictionary itself grows with new additions and the annotations used in the corpus become more accurate. This leads to a systematic evolution of both the dictionary and the corpus, effectively enhancing each of them.

## 2. Evaluation of Sign Language Resources

Although the STS Dictionary has been available online since 2000, it has never been evaluated. To assess the use of Swedish Sign Language resources, a questionnaire was sent to both old and new users. A web-based survey tool called Survey, commonly used for course evaluations at Stockholm University, was used for this purpose. This interface is versatile and suitable for various types of surveys. In this instance, it was used to evaluate the use of Swedish Sign Language resources. A link was provided to a signed version of the questionnaire containing 25 questions. The survey was conducted in Sweden over a two-week period in January 2024. The questionnaire included both open-ended and closed-ended questions. A total of 249 responses were collected from Sweden’s three regions: 7.2 percent from Norrland, the northernmost region; 56.2 percent from Svealand in south-central Sweden; and 34.5 percent from Götaland in southern Sweden (see Figure 3).

Demographically, 43 percent of the respondents identified as deaf, 14.5 percent as hard-of-hearing, 1.6 percent as late-deafened, 0.8 percent as deaf-blind, 4.4 percent as children of deaf adults (CODA),

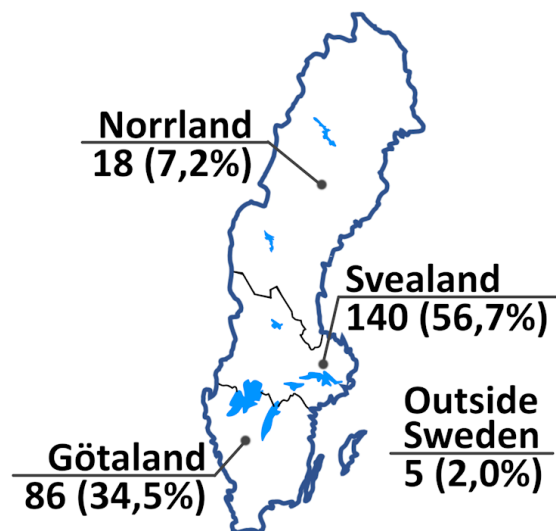


Figure 3: Percentage of responses from each of Sweden’s three regions.

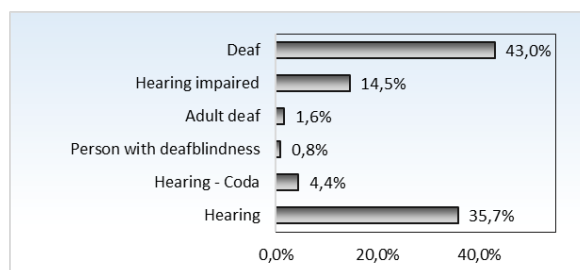


Figure 4: Demographic data on the deaf/blind/hearing status of the respondents.

and 35.7 percent as hearing (see Figure 4). The majority reported being born in the 1960s, 1970s or 1980s, with these three decades accounting for 73 percent of respondents. The remaining respondents were born in the 1950s or 1990s, with approximately 7–10 percent falling in each of these categories (Figure 5).

Among the respondents, 47 percent stated that Swedish Sign Language (STS) was their first language, and nearly 47 percent that Swedish was their first language. Three percent stated that some

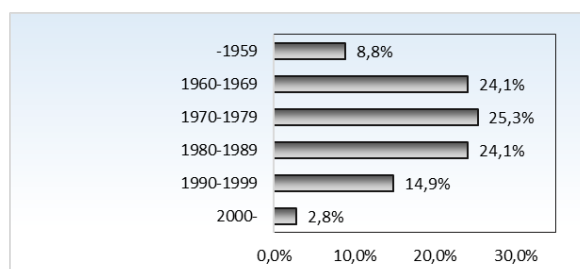


Figure 5: Demographic data on the year of birth of the respondents.

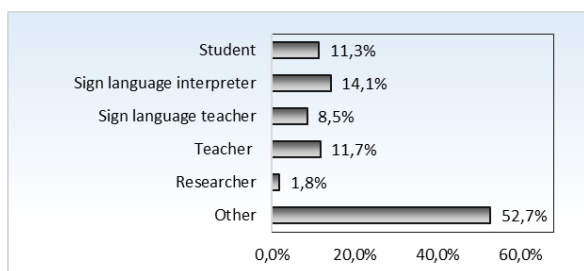


Figure 6: Demographic data on the occupation of the respondents.

other signed language was their first language. Regarding competence in STS, 80 percent of the 249 respondents assessed their proficiency as high or relatively high. In terms of occupation, 12.9 percent of respondents were students, 16.1 percent sign language interpreters and 22.8 percent educators/teachers/instructors at levels varying from preschool to university. Only two percent were researchers. See Figure 6).

### 3. Results

The popularity of online sign language resources has been evident for many years. However, the survey provides valuable insights into users' views on STS resources. The results are divided into the following categories:

- Facebook group
- Use of language resources
- Frequency of use of the STS Dictionary and STS Corpus
- Searching for signs/words, topics, functions, tools
- Opinions about resources (grammar, function, manual)

#### 3.1. Facebook Group

The Facebook group *Teckenspråkslexikon* [Sign Language Dictionary] was established in 2014 and is managed by the STS Dictionary team. It currently boasts some 6,998 members and continues to gain in popularity, with 93.2 percent of the 249 respondents stating that they are a member of the group. The survey conducted for this paper included both closed- and open-ended questions regarding respondents' engagement with the group. In response to an open-ended question asking what they liked about the group, 182 respondents provided brief answers. Positive feedback included comments such as "good", "excellent", "fascinating discussions", "many new signs" and "have learned

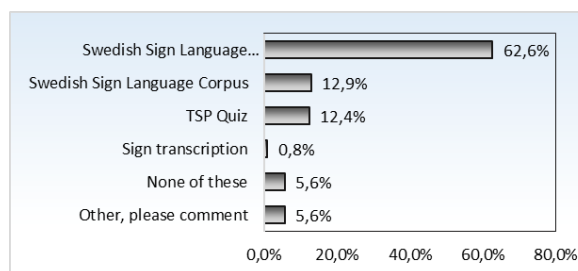


Figure 7: Respondents' use of sign language resources.

new signs". There was also, some critical feedback, with comments such as "good, but it is hard to follow those who sign fast", "want to know about sign semantics", and "why do many write when it is possible to sign". Overall, the Facebook group serves as an important platform for individuals to engage in various discussions and exchange information related to sign language usage, lexical variation and sign formation.

#### 3.2. Use of Language Resources

Statistics on the use of all or selected STS resources are not available. However, traffic statistics for the STS Dictionary and *STS-korpus* show the number of visitors to the websites. Between January 2023 and January 2024, there were over four million visits and 93,977 unique visitors. Two peak months for visits are February and October (1–2 months after course start). The survey provides insights into the use of specific STS resources (Figure 7). According to the 249 respondents, 90.7 percent use STS Dictionary, while 18.7 percent use *STS-korpus*. Additionally, 17.9 percent use the TSP Quiz [tspquiz.se](http://tspquiz.se), primarily for self-learning purposes, while 1.2 percent use the sign transcription tool. Other resources not administered by Stockholm University include *Spread The Sign* and *Teckenspråkspedagogerna* [Sign Teachers], which 8.1 percent of respondents use. Notably, 8.1 percent of respondents chose not to disclose this information. Most respondents learned of these resources via the internet (42 respondents), Facebook (39 respondents), educational settings (38 respondents), friends (15 respondents) or colleagues (13 respondents).

#### 3.3. Frequency of Use of Language Resources

The survey collected data on the average frequency of everyday use of sign language resources. When asked "How often do you use the Swedish Sign Language Dictionary?", 13.5 percent of the 249 respondents replied that they use the STS Dictionary on a daily basis, 18.8 percent that they use it

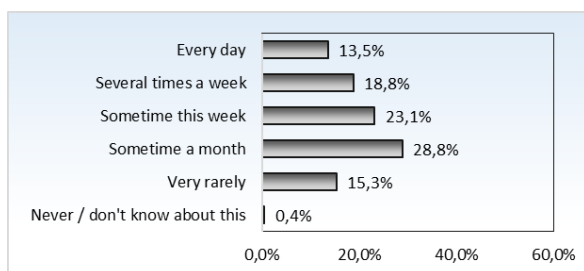


Figure 8: Respondents on how often they use the STS Dictionary.

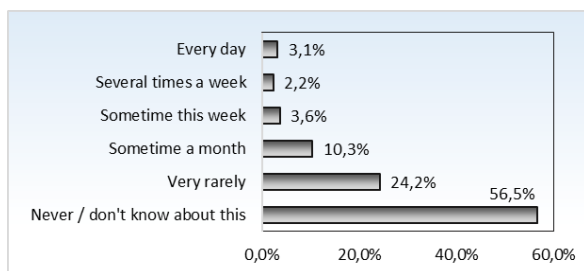


Figure 9: Respondents on how often they use of the STS Corpus.

several times a week, 23.1 percent that they use it at some point during the week and 28.8 percent that they use it once a month (Figure 8).

Regarding the STS Corpus, many respondents indicated that they were unfamiliar with it; specifically, 24.2 percent reported using it very rarely and 56.5 percent that they never used it or had not heard about it (Figure 9). Two main reasons were cited for not using it: some users only need to find a sign and do not require analysis of linguistic constructions in conversations or narratives.

With regard to the question "How do you feel about using the Swedish Sign Language Dictionary?", the parameters user-friendliness, relevance, comprehensibility and effectiveness in aiding language learning or communication were rated on a scale ranging from "easy" = 0 (left) to "hard" = 10 (right). The scale is wide at both ends, with ratings of 1–3 indicating ease (easy) and 8–10 indicating difficulty (hard). See Figure 10 for a graphical representation.

Additionally, responding to open-ended questions, fifteen respondents commented that they found the STS Dictionary to be clear, user-friendly and easy to navigate. They expressed appreciation for the explanations provided in STS through videos. Six respondents occasionally find navigating the system challenging and request the return of the previous "Slow down" function to help them keep pace. Additionally, five respondents find it difficult due to the lack of words and uncertainties about which sign variants are predominantly used, especially with limited knowledge of STS. They also

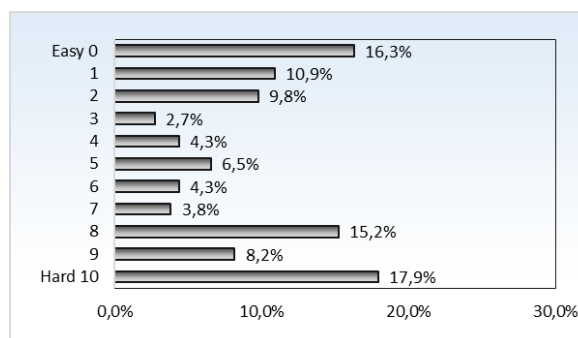


Figure 10: Respondents on the user-friendliness of the STS Dictionary.

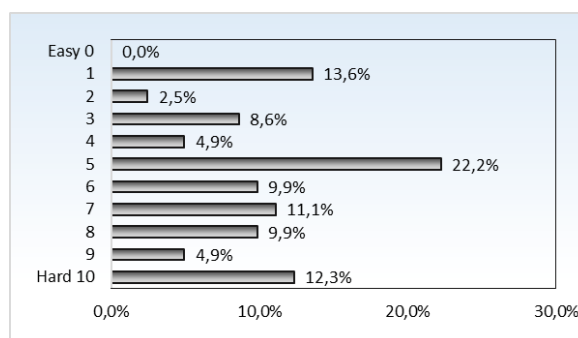


Figure 11: Respondents on the user-friendliness of the STS Corpus.

express confusion about which regional and social sign variants should be used. The question regarding the user-friendliness of *STS-korpus* is posed similarly: "How do you feel about using *STS-korpus*?". On the scale from "easy" = 0 (left) to "hard" = 10 (right), responses to this question generally fall in the middle of the scale, as shown in Figure 11).

In addition, 25 respondents state that they do not use or are unfamiliar with *STS-korpus*. Four respondents consider *STS-korpus* to be user-friendly. Three respondents stated they would like to use it more frequently and require more information about its functions, such as a link between *STS-korpus* and STS Dictionary.

### 3.4. Searches

Various tools have been developed to facilitate sign searches in the STS Dictionary. These tools provide several different search paths, including:

- searching with Swedish words
- searching for signs
- exploring other meanings of signs
- identifying alternative signs
- filtering signs by handshake

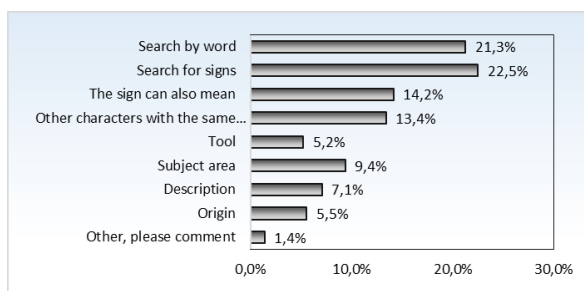


Figure 12: How respondents use search functions in the STS Dictionary.

- browsing signs by subject area
- searching with English words

Additionally, users can search for signs directly within the STS corpus, thus enhancing the comprehensibility and usability of the STS Dictionary. These features collectively improve the accessibility and effectiveness of sign language lookup and comprehension for users. Typing a Swedish word into a search box is the most common search procedure. The four most commonly searched words and phrases are "jag älskar dig" [I love you], "mamma" [mother], "hur mår du" [How are you?] and "jag heter" [My name is]. However, this does not necessarily imply that each sign corresponds directly to a word in Swedish. Instead, users search for signs that have an equivalent meaning in Swedish. Another common method of searching for signs is to enter subject categories and select a topic, such as healthcare, or a subcategory within a topic. One unique feature of the STS Dictionary is the ability to search by sign form, allowing users to specify particular characteristics of signs they are looking for.

The survey results confirmed that 64 percent of the 249 respondents searched for signs directly, while 60.5 percent searched using Swedish words. Additionally, 40.4 percent wanted to find signs with other meanings, and 38.2 percent searched for synonyms for signs they already knew. Furthermore, 26.8 percent of the 249 respondents searched for signs within specific subject areas, while 20.2 percent searched for descriptions of signs and 15.8 percent were interested in etymology. Finally, 14.9 percent selected tools for searching in various ways. (See Figure 12) for a graphical representation.)

### 3.5. Opinions about STS Resources

When asked "Do you use the manual for each language resource?", 61 percent of the 249 respondents replied that they were unaware that there were manuals available for the STS resources, or where they could find written and signed instructions on using the STS Dictionary and the STS

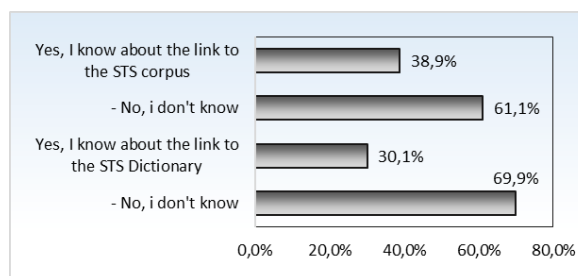


Figure 13: Respondents' awareness of links between the STS Corpus and Dictionary.

Corpus. On the next question, "Do you know that you can click on a place next to a sign entry in the STS Dictionary to see its use in the STS Corpus?" 61.1 percent were unaware of the link from the dictionary to the corpus and 69.9 percent were unaware of the link from the corpus to the dictionary (Figure 13).

There is divided opinion on whether the STS Dictionary website should include grammar or if grammar should be available on a separate website: 42.6 percent of the 249 respondents believe it is beneficial to integrate reference grammar into the dictionary, 13.5 percent would prefer reference grammar to be on separate pages, while 43.9 percent have no opinion on the matter.

Ultimately, the decision depends on factors such as user preferences, usability, and the overall design and functionality of the dictionary. Respondents were asked: "Do you think the Swedish Sign Language Dictionary should include grammar, or should grammar be on a separate website?" It is not easy to evaluate the response to this question, even among respondents who are sign language teachers, instructors or educators (only 22.8 percent of all respondents) or whether they have been informed about reference grammar. However, the survey results show that integrating grammar directly into the STS Dictionary might provide convenience for users who prefer having everything in one place, while having it on separate pages could offer a more organized and focused approach to learning grammar.

The tool *STS-korpus* retrieves material from the corpora and is used for the expanded presentation of how signs occur in natural language usage. Additionally, the survey highlights a need for grammar description and other functions, indicating the importance of considering user feedback and incorporating necessary features to enhance the usability and effectiveness of the resources. A reference grammar is a planned project.



## 4. Conclusion

The development work is intended to make the web-based dictionary more user-friendly and useful across different platforms. STS-korpus is an online interface for the Swedish Sign Language Corpus linked to the STS Dictionary, enhancing its utility. The extensive experience and language skills of the dictionary team are indispensable in the work of developing, updating, and improving the dictionary. The survey results confirm that work on STS language resources is on the right track.

The results of the survey concerning how the resource are perceived by users are both expected and unexpected, but valuable nonetheless.

- The Facebook group is much-appreciated as a forum for discussion and asking for specific signs in the signing community.
- Unsurprisingly, the STS Dictionary is the most used resource. The survey provides insights into the use of specific STS resources and how often people use them.
- Respondents describe how they search for signs/words, topics, features, and tools, which helps the team develop the features in the STS language resources.
- The survey also reveals varying opinions on ease of use, search functions, manuals and integrated grammar. While the results confirm what was suspected and anticipated, it is useful to have objective and quantifiable confirmation. The responses will guide future work on the language resources, including improving instructions and developing an interface linking the STS Dictionary and STS Corpus with a planned STS grammar resource.

## 5. Bibliographical References

Brita Bergman. 1979. *Signed Swedish*. Liber, Stockholm.

Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, Eleni Efthimiou, Neil Fox, Onno Crasborn, Lianne Westenberg, Sarah Ebling, Laure Wawrinka, Johanna Mesch, Thomas Björkstrand, Anna Kuder, and Joanna Wójcicka. 2024. [Extended interlingual index for the project's core sign languages and languages covered in WP9](#). Project Note D6.5, EASIER Consortium, Hamburg, Germany.

Thomas Björkstrand, Eira Balkstam, and Josephine Willing. 2022. [Svenskt teckenspråkslexikon som ett digitalt verktyg i andraspråksundervisningen: Poängberäkning av meningar \[Swedish Sign Language Dictionary as a digital tool in second language instruction: Point calculation of sentences\]](#). Report, Department of Linguistics, Stockholm University, Stockholm.

Danielle Bragg, Kyle Rector, and Richard E. Lerner. 2015. [A user-powered american sign language dictionary](#). In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, page 1837–1848, New York, NY, USA. Association for Computing Machinery.

Trevor Johnston. 2008. [Corpus linguistics and signed languages: no lemmata, no corpus](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 82–87, Marrakech, Morocco. European Language Resources Association (ELRA).

Lorraine Leeson, Jordan Fenlon, Johanna Mesch, Carmel Grehan, and Sarah Sheridan. 2019. [The uses of corpora in L1 and L2/Ln sign language pedagogy](#). In Russell S. Rosen, editor, *The Routledge Handbook of Sign Language Pedagogy*, pages 339–352. Routledge.

Rachel McKee and Mireille Vale. 2023. [Recent lexical expansion in New Zealand Sign Language: context, scope and mechanisms](#). *Current Issues in Language Planning*, pages 1–22.

Johanna Mesch. 2023. [Creating a multifaceted corpus of Swedish Sign Language](#). In Ella Wehrmeyer, editor, *Advances in Sign Language Corpus Linguistics*, chapter 9, pages 242–261. John Benjamins Publishing Company, Amsterdam.

Johanna Mesch, Eir Elisabet Cortes, Thomas Björkstrand, Nikolaus Riemer Kankkonen, Joel Bäckström, and Patrick Hansson. 2023. [Teckenspråkslexikografi – utmaningar i en annan modalitet. \[Sign language lexicography - challenges in a different modality\]](#). In *Nordiska studier i lexikografi 16: Rapport från 16:e konferensen om lexikografi i Norden, Lund 27–29 april 2022*, pages 225–240, Lund/Göteborg. Skrifter utgivna av Nordiska föreningen för lexikografi. Skrift nr 17.

Johanna Mesch, Lars Wallin, and Thomas Björkstrand. 2012a. [Sign language resources in Sweden: Dictionary and corpus](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages

127–130, Istanbul, Turkey. European Language Resources Association (ELRA).

Johanna Mesch, Lars Wallin, Anna-Lena Nilsson, and Brita Bergman. 2012b. [Dataset. Swedish Sign Language Corpus project 2009–2011 \(version 1\)](#).

Zrajm Öqvist, Nikolaus Riemer Kankkonen, and Johanna Mesch. 2020. [STS-korpus: A sign language web corpus tool for teaching and public use](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 177–180, Marseille, France. European Language Resources Association (ELRA).

Nikolaus Riemer Kankkonen, Thomas Björkstrand, Johanna Mesch, and Carl Börstell. 2018. [Crowdsourcing for the Swedish Sign Language dictionary](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 171–176, Miyazaki, Japan. European Language Resources Association (ELRA).

Adam Schembri and Kearsy Cormier. 2022. Signed language corpora: Future directions. In Jordan Fenlon and Julie A. Hochgesang, editors, *Signed Language Corpora*, pages 196–220. Gallaudet University Press.

Lucia Vlášková and Hana Strachoňová. 2021. [Leksikografija znakovnega jezika: študija primera spletnega slovarja](#). *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 9(1):90–122.

## 6. Language Resource References

Svenskt teckenspråkslexikon. 2024. [Swedish Sign Language Dictionary Online](#). Sign Language Section, Department of Linguistics, Stockholms University. PID <https://teckensprakslexikon.su.se>.

Svensk teckenspråkskorpus. 2024. [STS-korpus \[web-based corpus tool\]](#). Sign Language Section, Department of Linguistics, Stockholm University. PID <https://teckensprakskorpus.su.se>.

# Sign Language Translation with Gloss Pair Encoding

Taro Miyazaki<sup>1</sup> , Sihan Tan<sup>1,2</sup> , Tsubasa Uchida<sup>1</sup> , Hiroyuki Kaneko<sup>1</sup> 

<sup>1</sup> NHK Science and Technology Research Laboratories

<sup>2</sup> Tokyo Institute of Technology

{miyazaki.t-jw, uchida.t-fi, kaneko.h-dk}@nhk.or.jp,

tan.s.ae@ra.sc.e.titech.ac.jp

## Abstract

Because sign languages are the first language for those who are born deaf or who lost their hearing in early childhood, it is better to use sign languages rather than transcribed spoken language to provide important information to these people. We have been developing a sign language computer graphics generation system to provide information to deaf people, and in this paper, we present a translation method from spoken language to sign language that can be used in the system. In general, since the number of glosses used when transcribing sign language is limited, a single meaning is often expressed by a combination of multiple sign words, i.e., the word “library” is expressed in Japanese Sign Language with two words: “book” and “building.” To merge these expressions into one token, we propose gloss pair encoding (GPE), which is inspired by byte pair encoding (BPE). This technique is expected to enable more accurate handling of expressions that have a single meaning in multiple sign words. We also show that it is effective as data augmentation on the sign language side in sign language translation, which has not been done much so far.

**Keywords:** Sign Language Translation, Machine Translation, Byte Pair Encoding, Gloss Pair Encoding

## 1. Introduction

Sign languages are typically the first language for those who are born deaf or who lose their hearing in early childhood. To provide information for these individuals, it is better to use sign language than to transcribe spoken language, as reading transcriptions of a spoken language, which is their second language, places an unnecessary burden on them.

We have been developing a sign language computer graphics (CG) generation system to provide information to deaf people. This system consists of two parts: a machine translation part and a CG generation part. The first part translates the input sentence of spoken language into a sign language gloss sequence, and the next part generates sign language CG based on the gloss sequence by referring to the motion data of each sign word. In this paper, we focus on improving the performance of the machine translation part.

In general, the number of glosses used for transcribing sign language is smaller than vocabulary size of spoken languages. For example, in our corpus, the gloss-based vocabulary size of Japanese Sign Language is approximately 4,000, while the word-based vocabulary size of Japanese is 27,000. Therefore, a single meaning is often expressed by a combination of multiple sign words, i.e., the word “library” is expressed in Japanese Sign Language with two words: “book” and “building.” In this case, the glosses “book” and “building” play the role of subwords. Also, some glosses play the role of a letter, as in fingerspelling. In other words, glosses are sometimes used as a word, sometimes as a subword, and sometimes even as a letter. Since

the granularity of glosses themselves can differ significantly, we believe that using them as they are in machine translation may cause degradation of the translation performance. Therefore, we propose a method to combine multiple glosses into one merged-gloss, and match the granularities.

The proposed method is named gloss pair encoding (GPE), which is inspired by byte pair encoding (BPE) (Sennrich et al., 2016). BPE is often used in machine translation to merge byte pairs that appear consecutively with high frequency into one merged subword. Our GPE is also merge gloss pairs that appear consecutively with high frequency into a merged-gloss. This allows multiple sign words that express one meaning to be treated as a single token. For example, the two glosses “book” and “building,” which express the meaning of library, can be treated as a single merged-gloss “book+building.” This is expected to enable more accurate handling of glosses with multiple meanings. Furthermore, through experiments, we demonstrate that it is also effective as data augmentation on the sign language side in sign language translation, which has not been done much so far.

Our contributions are summarized as follows. (1) We propose gloss pair encoding (GPE) to treat a gloss sequence that appears consecutively with high frequency as one token. (2) We show that by setting an appropriate number of vocabulary words, using merged-gloss with GPE can improve translation performance. (3) We experimentally demonstrate that by using a corpus with and without GPE in combination, translation performance can be improved due to the effect of data augmentation.

## 2. Related Work

### 2.1. Tokenization in Machine Translation

Machine translation utilizing neural networks originally treated words as the smallest unit (Cho et al., 2014; Bahdanau et al., 2015). Later on, in order to take advantage of the fact that each part of a word has something in common (e.g., the “person” part is common between personal and person, and there are also similarities in meaning) subwords have come to be used.

Byte pair encoding (BPE) is one of the most commonly utilized subword extracting methods. BPE was first proposed by Gage (1994) for encoding strings of text into tabular form, and Sennrich et al. (2016) then applied it to natural language processing methods including machine translation.

To avoid creating subwords that cross word boundaries, it is necessary to provide a separation for each word in advance. This is simple enough for languages that already have white spaces to separate words, such as English and German, but for other languages, such as Chinese and Japanese, it is necessary to separate words in advance. In response to this challenge, sentencepiece (Kudo and Richardson, 2018) was proposed as a method that allows end-to-end tokenization even in languages without word breaks.

### 2.2. Machine Translation of Sign Language

Several methods for translating spoken language into sign language have been proposed. Zhang and Duh (2021) regarded sign language translation as a low-resource machine translation task, and applied some of the techniques that are often used in low-resource language translation such as hyperparameter search and back translation. Zhu et al. (2023) applied techniques common to low-resource machine translation to sign language machine translation and showed that these techniques can also improve sign language translation. All of these methods use gloss sequence as sign language transcription.

The disadvantage of using gloss is that it causes important information in sign language, such as facial expressions and finger movement speed, to be lost. Therefore, gloss-free translation methods have recently proposed. Lin et al. (2023) proposed an end-to-end gloss-free translation method. Zhou et al. (2023) developed a novel pre-trained paradigm that combines masked self-supervised learning with visual language supervision learning, and they reported that this approach can deliver good translation. While gloss-free translation methods are currently used for translating sign language video into spoken language, there are few exam-

ples of its application for translating spoken language into sign language. This is because it is very challenging to generate motion data of sign language directly, which is necessary for gloss-free translation from spoken language to sign language,

The Conference on Machine Translation (WMT), a well-known workshop series of machine translation, initiated a shared task on sign language translation in 2022 (Müller et al., 2022). We hope this will lead to even more active research into sign language translation.

## 3. Proposed Method

As mentioned in the Introduction, the granularity of glosses can differ significantly, which is one of the reasons machine translation of sign language is difficult. Therefore, in our approach, we merge frequently occurring gloss sequences into one token by using gloss pair encoding (GPE), which is based on byte pair encoding (BPE) (Sennrich et al., 2016) and modified for sign language, to match the granularity of tokens.

First, we explain the original BPE, and next we present our proposed GPE.

### 3.1. Byte Pair Encoding (BPE)

BPE first initializes the vocabulary while covering all the characters in the training data, and regarding input data as sequences of characters with a special end-of-word symbol “.”, which is utilized to restore the subword segmentation sentence to the original sentence. It then counts the frequencies of all symbol pairs and replaces the most frequently used pair (“A”, “B”) into one merged-character ‘AB,’ and adds it to the vocabulary. BPE applies this process repeatedly until finally it outputs the vocabulary including all the characters and merged-characters. The final vocabulary size can be controlled as a hyperparameter of the number times to repeat merge operations.

BPE makes it possible to achieve subword segmentation, where sequences of characters with meaning becomes a single vocabulary.

### 3.2. Gloss Pair Encoding (GPE)

In our GPE, the operation is almost the same as with BPE but differs in that it compresses frequent pairs of glosses instead of frequent pairs of bytes.

GPE first initializes the vocabulary while covering all the glosses in the training data. Unlike BPE, GPE does not use a special end-of-word symbol “.”. It then counts the frequencies of gloss pair, and replace the most frequently used pair (“gloss<sub>A</sub>”, “gloss<sub>B</sub>”) into one merged-gloss “gloss<sub>A</sub>+gloss<sub>B</sub>,” and adds it to the vocabulary. GPE applies this



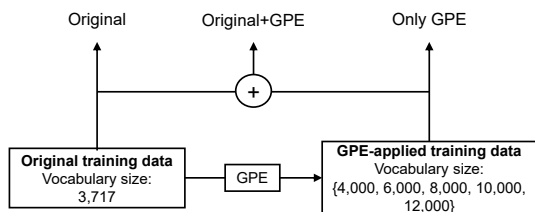


Figure 1: Three types of training data used in experiments.

process repeatedly until finally it outputs the vocabulary including all the glosses and merged-glosses. The final merged-gloss vocabulary size can be controlled as a hyperparameter of the number of times to repeat merge operations.

GPE merges gloss-pairs that appear continuously and frequently into a single merged-token. As a result, glosses that have the role of a subword are merged into merged-gloss, and we can expect the granularity of glosses to be ensured.

Also, by using GPE, the vocabulary size can be freely set using a hyperparameter, so the difference in vocabulary between spoken language and sign language can be reduced. This may also improve the translation performance.

### 3.3. Applying GPE to NMMs

Sign language utilizes non-manual movements (NMMs) such as head-nods and pointing. Note that although pointing is often treated as a sign word, we treated it as one of NMMs in this paper. Pointing does not express any meaning by itself, but it expresses meaning when combined with other words. Since this is the same as a head-nod, we decided to treat pointing as one of NMMs as a head-nod.

Head-nods often serve as function words, such as indicating the beginning or end of a sentence and expressing breaks in sentences or parallel relationships. Pointings is used to express meaning by referencing a previous word to emphasize the subject (Liddell, 2003).

In sign languages, NMMs are used much more frequently than sign words. Therefore, if NMMs are included in the merge target of GPE, it can be expected that many merged-glosses containing NMMs will be created. To evaluate this effect, we compare the performance when merging NMMs with GPE (with-NMM-GPE) and when not (without-NMM-GPE). In without-NMM-GPE, we first divide sentences by NMMs, then remove the NMMs, and finally apply GPE.

## 4. Experiments

We conducted an experiment to evaluate the text-to-gloss translation performance utilizing data with

different sizes of vocabulary using GPE. In the experiment, we prepared one setting that did not apply GPE (Original), one that used only training data that applied GPE (OnlyGPE), and one that both applied and did not apply GPE are merged as training data (Original+GPE), as shown in Figure 1. In Original+GPE, training data with two different vocabularies are mixed and shuffled for learning. Since the vocabulary expanded by applying GPE always includes the same vocabulary as Original, it is possible to learn with a single encoder-decoder model without having to separate the translation models. We did not apply GPE to the development and test data, and the merged-gloss of the translation result was restored to the original gloss for evaluation.

In the experiment, we prepared training data with a total of five patterns of vocabulary size by applying GPE: 4,000, 6,000, 8,000, 10,000, and 12,000 for both with-NMM-GPE and without-NMM-GPE. As a baseline, we also conducted an experiment in which the vocabulary size of 3,717 that appeared three or more times in the corpus was used without GPE. The number of Japanese vocabularies was set to 8,000 in all experiments. We describe the experiments in detail below.

### 4.1. Our Corpus

We used an in-house corpus called the Japanese-JSL sign language news corpus for our experiment. This corpus is created from daily NHK sign language news programs, which are broadcast on NHK TV with Japanese narrations and JSL signings. The corpus includes around 160,000 Japanese transcriptions, JSL transcriptions, and JSL videos. Japanese is transcribed by revising the results of applying speech recognition on news programs. JSL is transcribed by native signers who manually transcribe each sign motion into sign language gloss. Note that, Japanese and JSL sentence pairs are not literal translations, so there are many subject complements, omissions, and so on. We transcribed all of the manual and some of the NMMs (e.g. head-nods and pointing) in linear transcription. In most cases, these type of manual and non-manual features are not expressed at the same time, so this transcription simplifies the JSL expressions while simultaneously retaining most of the necessary information.

We selected 129,950 sentence pairs that do not include *classifier*, which is hard to be transcribed into gloss. This is because *classifier* has a large vocabulary and no fixed hand or finger expressions, so our sign language CG generation system cannot convert them into sign language CG. We randomly split the corpus into 127,950 for training, 1,000 for development, and 1,000 for testing.

Data	No. of vocab.	without-NMM-GPE		with-NMM-GPE	
		Median	Average & std. deviation	Median	Average & std. deviation
Original	3,717	24.27	24.46 ± 0.40	–	–
Only GPE	4,000	<u>24.75</u>	<u>24.73</u> ± 0.20	<u>24.53</u>	24.40 ± 0.47
	6,000	24.69	24.81 ± 0.14	21.75	21.74 ± 0.15
	8,000	24.09	24.05 ± 0.07	21.49	21.42 ± 0.17
	10,000	23.61	23.72 ± 0.35	20.83	20.72 ± 0.26
	12,000	23.45	23.46 ± 0.29	21.00	20.87 ± 0.23
Original + GPE	4,000	<u>24.36</u>	23.74 ± 0.60	<u>24.37</u>	24.37 ± 0.05
	6,000	<b>25.03</b>	25.03 ± 0.05	<u>24.94</u>	24.74 ± 0.47
	8,000	24.79	24.81 ± 0.09	24.75	24.66 ± 0.26
	10,000	<u>25.02</u>	<b>25.05</b> ± 0.09	<u>24.64</u>	24.59 ± 0.16
	12,000	<u>24.61</u>	24.79 ± 0.35	<u>24.75</u>	24.58 ± 0.51

Table 1: Experimental results. We show the median, average, and standard deviation of BLEU after three attempts with different random seeds. **Bold** indicates the best result in the table, and underline indicates a result that outperformed original.

## 4.2. Experimental Setting

We utilized a 6-layer transformer encoder-decoder model (Vaswani et al., 2017) with the *norm-first* setting (Nguyen and Salazar, 2019) for the translation model. We utilized PyTorch (Paszke et al., 2019) for implementing the model and RAdam (Liu et al., 2020) for optimization with the learning rate of  $1.0 \times 10^{-3}$ . We utilized cross-entropy loss for calculating loss in training. The dropout ratio for the transformer encoder and decoder was 0.1, and that for the output layer of the feed-forward neural network was 0.3. We applied sentencepiece (Kudo and Richardson, 2018) for input Japanese sentences with a vocabulary size of 8,000. We trained the models with the batch size of 256 and the number of training epoch of 50. We evaluated the model in each epoch using the development data, and chose the model with the best BLEU score on development set. We trained the models three times with different random seeds.

### 4.2.1. Results

Experimental results are provided in Table 1. As shown, Original+GPE performed better than Original in a wide range of vocabulary sizes for both without-NMM-GPE and with-NMM-GPE settings. In contrast, OnlyGPE performed better only in small vocabulary size, and its performance was worse than Original+GPE. In particular, OnlyGPE using the with-NMM-GPE setting underperformed the baseline in most cases. Overall, Original+GPE using the without-NMM-GPE setting performed the best.

## 4.3. Discussion

### 4.3.1. with- and without-NMM-GPE

The without-NMM-GPE setting outperformed the with-NMM-GPE setting for almost all vocabulary

No. of vocab.	%
4,000	78.5
6,000	66.1
8,000	60.6
10,000	58.4
12,000	57.5

Table 2: Percentage of merged-gloss including NMMs in with-NMM-GPE setting.

Merged-gloss	Meanings
without-NMM-GPE	
<i>Explanation + Disappear</i>	I have given an explanation
<i>Decide + Disappear</i>	It has been decided
<i>Place + Place</i>	In various places
<i>People + Everyone</i>	Everyone
<i>High + Temperature</i>	Highest temperature
with-NMM-GPE	
<u>pointing + head-nod</u>	–
<u>Exist + head-nod</u>	Existing (EOS)
<u>Disappear + head-nod</u>	Finished (EOS)
<u>head-nod + pointing</u>	–
<u>In + head-nod</u>	Still (EOS)

Table 3: Top-5 merged-gloss of with-NMM-GPE and without-NMM-GPE. Glosses are *italic*, NMMs are underline, and merge is denoted by “+”. (EOS) indicates that the head-nod in merged-gloss marks the end of a sentence.

sizes. NMMs are very often used in sign language, so if GPE merges NMMs, most of the merged-glosses contain NMMs, and sign words are less often merged. Table 2 gives the percentage of merged-glosses that contains at least one NMMs, and Table 3 shows the top-5 frequently appearing merged-glosses in training data for the with and without-NMM-GPE settings. More than half of the merged-glosses in with-NMM-GPE contain NMMs, and many merged-glosses are combined with a head-nod representing the end of a sentence.

Data	# vocab.	% of merged-gloss	
		Train	Output
Original	3,717	0	0
	4,000	0.35	3.1
Only	6,000	12.91	11.65
GPE	8,000	17.52	14.60
	10,000	20.00	15.82
	12,000	21.66	16.66
	4,000	0.18	0.01
Original	6,000	6.01	0.50
+ GPE	8,000	7.94	0.55
	10,000	8.92	0.86
	12,000	9.56	0.88

Table 4: Percentage of merged-gloss among all gloss.

Since this combination does not extend the meaning, it is presumably not very effective. In contrast, many merged-glosses of without-NMM-GPE take on new meaning by merging multiple glosses.

Head-nods, which are a type of NMMs, often serve as function words and do not express meaning when combined with other words. Therefore, merging head-nods using GPE does not seem very effective. In contrast, pointing expresses meaning by combining with the previous word. However, our GPE often merged a pointing with the word following the pointing, which is meaningless since the pointing always points in the direction of the previous word. Also, as reported in our previous research, machine translation that merges a pointing and the previous word does not improve the performance (Miyazaki et al., 2020), thus, demonstrating that merging NMMs did not contribute to improving the translation quality.

We found that a better performance could be obtained by excluding NMMs from the merge target in GPE. In the following discussion, we examine the case of using without-NMM-GPE.

#### 4.3.2. Effects of Merged-gloss

In the OnlyGPE setting, when the number of vocabularies was set appropriately (i.e., vocabulary size of 4,000 or 6,000 in this experiment) the performance improved. This indicates that with an appropriate vocabulary size, expressions that express one meaning by using multiple glosses can be combined into a merged-gloss, which makes it easier for translation models to learn and thereby improves the performance.

In contrast, OnlyGPE does not improve when the vocabulary size is increased to the same level as Japanese. This shows that it is not necessary to match the number of vocabularies in the source and target languages. The performance deterioration of OnlyGPE when the vocabulary size is large is presumably due to the fact that gloss-pairs

that are not very frequent in the training data were also merged. The percentage of merged-glosses among all glosses for training data and translation output is shown in Table 4. With OnlyGPE, as the number of vocabularies increases, the percentage of merged-glosses in the output becomes considerably smaller compared to the training data. This suggests that GPE can create merged-glosses that are actually useful in translation only when the number of vocabulary words is around 6,000 in this dataset, and that for larger vocabulary sizes, it was mostly noise during learning. With OnlyGPE, merged glosses are not learned as a single gloss, so the influence of noise will be greater. On the other hand, with Original+GPE, merged glosses can be learned as merged-gloss as well as each single gloss by using original data, so the influence of noise can be reduced. This is why Original+GPE performed better than OnlyGPE especially for large vocabulary size in the experiments.

#### 4.3.3. Effects of Data Augmentation

As shown in Table 4, the outputs of Original+GPE include not so many merged-glosses compared with the percentage of merged-glosses in the training data. This suggests that the increase in the amount of training data—that is, the effect of data augmentation—was large in Original+GPE and had a greater influence than the effects of the merged-glosses. Data augmentation in gloss-based sign language translation has been reported by (Zhu et al., 2023) and while they demonstrated data augmentation due to differences of the pre-processing on the spoken language side, data augmentation for sign language side was not examined. Our experiments indicate that data augmentation on the sign language side is also effective.

## 5. Conclusion

In this paper, we presented a translation method using gloss pair encoding (GPE), which merges multiple consecutive sign words that frequently appear in a corpus. We experimentally demonstrated that the translation performance improved when applying GPE with an appropriate number of vocabularies. We also found that by learning together with a corpus to which GPE is not applied, the effects of data augmentation can be obtained and translation performance can be further improved. When applying GPE it is better not to merge NMMs such as head-nod and pointing.

We did not perform an experiment in combination with data augmentation on the spoken language side. Many data augmentation methods for spoken language have been proposed, so considering how to combine them will be left as our future work.

## 6. Limitation

Also, this time we only used the training dataset of our in-house Japanese-Japanese Sign Language corpus. We hope we are confident that performance will improve regardless of the language pair, but we have not yet conducted experiments with other languages.

## 7. Bibliographical References

- Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. [On the properties of neural machine translation: Encoder–decoder approaches](#). In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.
- Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Scott K Liddell. 2003. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press.
- Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. [Gloss-free end-to-end sign language translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.
- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. [On the variance of the adaptive learning rate and beyond](#). In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.
- Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. [Machine translation from spoken language to sign language using pre-trained language model as encoder](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).
- Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. [Findings of the first WMT shared task on sign language translation \(WMT-SLT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Toan Q. Nguyen and Julian Salazar. 2019. [Transformers without tears: Improving the normalization of self-attention](#). In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural*



*Information Processing Systems*, volume 30. Curran Associates, Inc.

Xuan Zhang and Kevin Duh. 2021. [Approaching sign language gloss translation as a low-resource machine translation task](#). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 60–70, Virtual. Association for Machine Translation in the Americas.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. [Neural machine translation methods for translating text to sign language glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

# SignCollect: A ‘Touchless’ Pipeline for Constructing Large-scale Sign Language Repositories

Gomèr Otterspeer<sup>1</sup>, Ulrika Klomp<sup>2</sup>, Floris Roelofsen<sup>3</sup>

g.otterspeer@uva.nl, u.klomp@uva.nl, f.roelofsen@uva.nl

## Abstract

The project team of the Signbank project at the University of Amsterdam intends to substantially extend the NGT lexicon in Global Signbank within a limited timespan. To make this possible, the signCollect platform was developed to automate a major part of the workflow. The signCollect system includes a ‘touchless’ interface which enables a signer to control the system through simple gestures (recognized using computer vision) to (i) prompt the display of the next lexical entry, (ii) start a new recording, and (iii) approve/disapprove a recording. This capability allows a signer to record between 60 to 120 signs per hour, without the need for any assisting staff to be present. The approved recordings immediately become visible in the signCollect database, so that other members of the team can add metadata. With feedback from workshop participants we intend to further optimize the signCollect platform and make it available as an open-source tool for all sign language research teams.

**Keywords:** Sign Language, Data Collection, Automated Data Harvesting

## 1. Introduction

### 1.1. The Signbank Project

The Signbank project at the University of Amsterdam aims at extending the NGT (Sign Language of the Netherlands) dataset in Global Signbank. Global Signbank is a lexical database utilized to collect, store and display signs from various sign languages around the world. Various research institutes employ Signbank for the publication and collection of lexical entries and use it to conduct field research (Cassidy et al., 2018). The aim of the current Signbank project is to substantially extend the NGT dataset (see Klomp et al., 2024 for more details on the general project) within a relatively short timespan of 14 months. In this project, the team aims to identify, record, and describe thousands of lexical signs. To automate parts of this process, the first author of this paper – a Deaf programmer and native signer of NGT – developed the signCollect platform.

### 1.2. Overview of the signCollect Platform

signCollect has three components:

1. signCollect Studio for recording;
2. signCollect Dashboard to support collaborative work among team members;
3. signCollect Hub which connects signCollect Studio and signCollect Dashboard to several other components of the larger software ecosystem in which signCollect operates.

The platform aims to facilitate discussions to identify signs to be added to the dataset, collecting

multi-view video recordings and 3D motion capture recordings of signs, managing validation of the recordings, and gathering metadata (phonological and semantic descriptions, discussion notes). All information related to a given sign is collected in one entry, with one Annotation ID gloss (following the structure of lexicons on Global Signbank). The platform streamlines the process of collecting data and aims to serve as a basis for possible future extensions (see Section 7) which may further enhance the data collection process. The NGT dataset on Global Signbank and the signCollect platform running at the University of Amsterdam are synchronised every time there is a change in their databases so that they always share the same dataset (see also Figure ??).

### 1.3. Other Approaches

Other projects working on sign language lexicons or corpora have also developed tools to support data collection and data management. Prominent examples are the GlossLexer system of Hanke et al. (2001) and the iLex system of Hanke and Storz (2008). In comparison to these existing systems, signCollect aims to further optimize the data collection and data management pipeline by automating several steps in the workflow, and also aims for smooth integration with the Global Signbank platform (although it may also be used in projects that do not make use of Global Signbank).

### 1.4. Paper Outline

The paper is organized as follows. First, Section 2 goes into the touchless recording procedure in signCollect Studio – the most innovative component of the signCollect platform. Then, Section 3 discusses

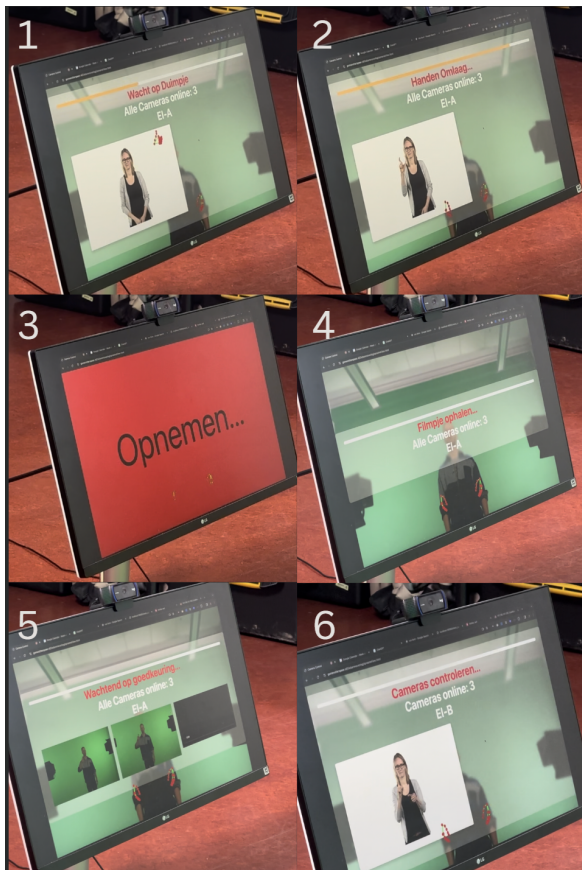


Figure 1: Touchless User Interface

the ecosystem that signCollect is part of, Section 4 turns to the general workflow that signCollect facilitates, Section 5 highlights some specific hardware and software components, Section 6 describes two concrete issues that signCollect helps to resolve, Section 7 discusses some avenues for future extensions, and Section 8 concludes.

## 2. A 'Touchless' Recording Pipeline

**Setup** Our basic recording setup involves three cameras. After placing and connecting the cameras and lights, and starting up the signCollect system, the system first checks if the cameras are connected and have the desired settings; if not, it will display a warning pop-up, which means that the settings have to be adjusted. When all settings are as desired, the warning pop-up disappears. The signer views a list of lexical entries that need to be recorded and also has the freedom to add/remove entries. The signer does a first set of recordings to check lights, their position w.r.t. the cameras, and to test that signCollect (which includes Mediapipe Gesture Recognition) correctly recognizes their control gestures (thumbs up and down). Then they can proceed. The steps described below are also visualized in Figure 1.

**Step 1: The signer views the lexical entry** The Annotation ID gloss of a lexical entry is being displayed, together with an earlier made quick recording so the form of the sign associated with this particular Annotation ID gloss is clear. When ready the signer gives a thumbs up for half a second. The interface shows a progress bar that fills to 100 percent as long as the thumb is up and no extreme movements are detected. This means that the progress can be stopped anytime by interrupting the thumbs up gesture.

**Step 2: The signer is ready to produce the sign** The application first checks again if the settings of all cameras are the same, the batteries are at a level of at least 10 percent and that there is sufficient space on the memory cards. Then it displays 'Ready to Record'. When seeing this, the signer lowers their hands. To avoid the signer looking at the display instead of at the front camera, the display turns bright red, which is a signal that is also visible from the corner of one's eyes.

**Step 3: Cameras start capturing** The application detects that the hands are down for at least half a second, then it proceeds to send a START CAPTURE signal to all cameras. The cameras send back a confirmation that they are recording. The system displays a message 'Recording' with a red background, and the cameras are rolling. Otherwise the background turns blue and a warning message is shown that the application hasn't received confirmations from all cameras; this can happen when one of the cameras has crashed or when the USB connection is interrupted. When the application has received all confirmations it also produces a high pitch sound for the purposes of synchronizing the video recordings of the different cameras. The signer can start producing the sign.

**Step 4: The sign has been captured** When the signer has produced the sign, the hands are lowered to a rest position. The system recognizes this and shows a progress bar as visual feedback for half a second, as long as there is no other extreme movement detected. The videos captured by the cameras are displayed. Depending on the settings of the user, the videos are either displayed all at once, or alternatively one after the other.

**Step 5: The signer approves or rejects the recordings** To approve the recordings, the signer makes a thumbs up gesture. Again a progress bar will be shown for one second, and then the captured videos are saved in a database to batch process them when the signer has finished recording all the listed signs. However, when the captured videos do not meet the requirements, the signer uses the

thumbs down gesture. The system shows the same lexical entry again to recapture the sign and ignores the rejected videos.

### Step 6: signCollect Studio collects all the data

When all the signs from the list are recorded, a window appears and says: 'ALL DONE'. Then the system asks the cameras to switch from Capture State to Media Transfer state. After that the videos are downloaded. A download progress bar is shown to the signer with a warning not to touch the USB-C cables and cameras. After all video files are downloaded from the cameras to the studio computer, they are uploaded to the signCollect database on a file server. The videos can immediately be viewed by other team members through a web interface. They can validate the video recordings and add metadata.

We now show how this capture procedure fits into the bigger ecosystem and how it relates to data storage and processing.

## 3. The signCollect Ecosystem

Figure ?? shows a diagram of the structure and data flow within the signCollect ecosystem. There are various components with connections between them – the following paragraphs explain the components from left to right.

**Global Signbank** Global Signbank and signCollect communicate with each other, receiving, sending, and acknowledging requests and responses. The communication uses standard JSON arrays to represent lexicon entries. When a user updates a field in signCollect, the platform saves the entry in the database. If the lexicon entry is ready to be published, signCollect sends the JSON array to Signbank. Signbank then updates the entry and replies with 'success' or 'error' to indicate whether the format adheres to the JSON standard and the relevant fields have been successfully updated.

**signCollect** signCollect is connected to all components of the ecosystem and is linked to a database on a separate server. The database contains information of all the lexical entries that are saved via the signCollect interface. Depending on the status of entries, connected components get a request from signCollect to process the metadata of the respective entry.<sup>1</sup>

---

<sup>1</sup>Every entry has fields for preliminary recordings, phonological information, gloss name, and media files. Depending on how many fields are entered and checked the platform defines a certain status for the entry.

**Studio hardware** The Signbank team in Amsterdam uses a video recording studio and a 3D Motion Capture studio. The hardware in these studios is controlled by signCollect, their output is processed and files with metadata are saved in the database on the file server.

**Data Storage** The storage contains video files, motion capture files, FBX/GLB files, a database, and metadata for AI tools. Every file is linked with a signed consent form of the participant involved. If there is a request to process the files, signCollect checks whether permission for publication has been given or not. The platform also regularly checks if consent has expired. In this case, the files are deleted (or in the future possibly anonymized using AI/ML with a human check).

**AI/ML Server** The server contains AI tools that are helpful for processing video files and motion capture files. In the future, additional AI tools will be installed to support the Signbank project team in adding metadata (e.g., automatic recognition of phonological features of a sign like the hand configuration or facial features) and also provides an environment for researchers and students to test new tools on the data (see Section 7).

## 4. The signCollect Workflow

### 4.1. Preparation

Figure 3 provides an overview of the signCollect workflow. During the data preparation phase, the Signbank team creates a list of signs to be added to the dataset. Then they need to make sure these signs do not already exist in the dataset. Therefore, phonological information is added as searchable metadata – a feature already available in Signbank. If the sign is not available in Signbank yet, we proceed by adding a lexical entry and a quick recording for easy reference using a webcam. See Klomp et al. (2024) for further discussion of the procedure of selecting signs.

### 4.2. Recording

When a substantial list of lexical entries with status 'ready-to-record' has been reached, a recording session is planned. The signer checks the list of lexical entries in the signCollect Dashboard before preparing the studio to record the signs. After preparing the studio with the required cameras and lights and checking the connections, the signCollect Studio system is started and checks the connections and properties of all cameras (battery and memory). If no warning is displayed on the screen of signCollect Studio, then the signer proceeds to



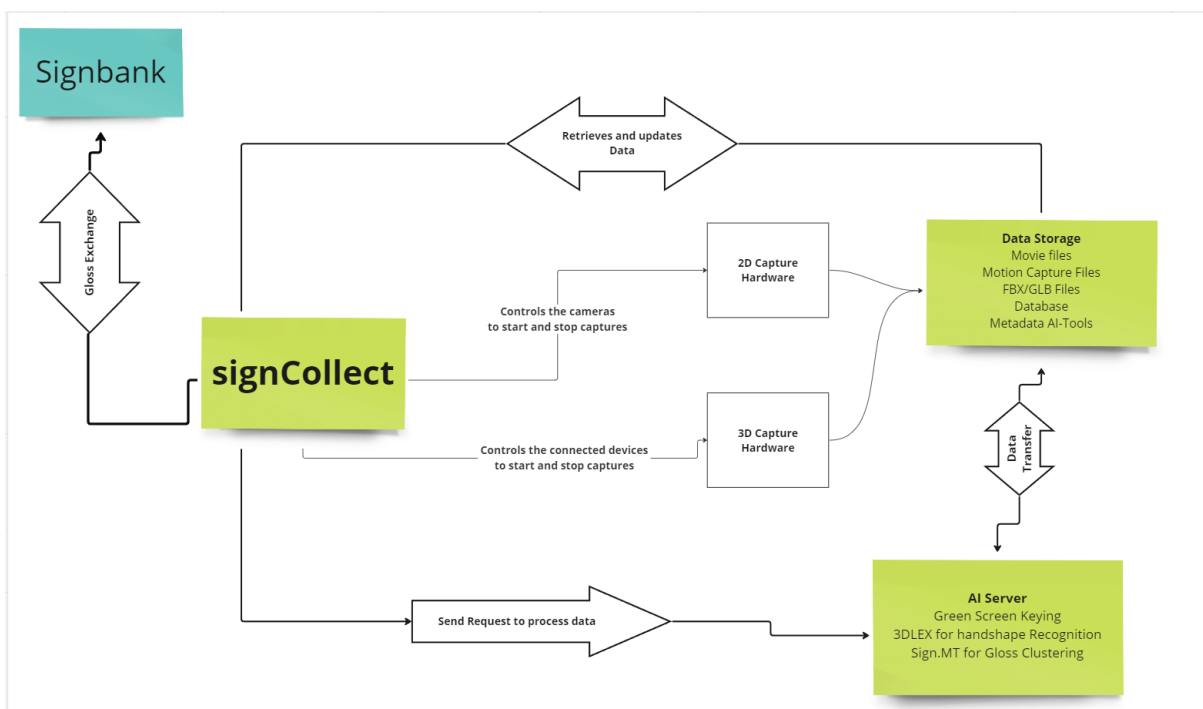


Figure 2: Intranet of signCollect



Figure 3: Workflow of signCollect

record by giving a thumbs up towards the main camera. The signer's control gestures are translated into commands by signCollect Studio (Section 5.2). Besides gesture control, signCollect Studio also offers signers another control option, making use of foot pedals. Both control options allow signers to work without interruption and without the need for additional technical assistance staff.

During each recording session, the system automatically verifies that data is stored properly and in the right format, making it immediately accessible to other team members. As described in Section 2, recordings are already reviewed during the recording session by the signer and re-recorded if needed to meet stringent quality standards, including a 60FPS frame rate, 4K resolution, absence of motion blur, proper lighting, and clear visibility of the sign. Recordings are captured from three distinct angles to provide comprehensive visual data. The same rigorous standards apply to 3D captures.

### 4.3. Automated Post Processing

In Automated Post Processing, signCollect Studio automatically retrieves files from the cameras and stores them on a file server. The signs are already

segmented because they are recorded individually. The file server is linked to an AI server, which compresses the files and performs keying, meaning that the green screen background is converted to a color determined by the Signbank project team. The AI server processes the recordings for Mediapipe Landmark Detection, which in future work could be used for recognizing phonological features such as hand configuration through various AI tools. This would further automate the workflow. In the end the recordings that have been made in the studio get the status 'To be checked before publication'. The Signbank project team reviews them and then approves or rejects them. When approved, the recordings get uploaded automatically to Global Signbank. When rejected, they go back to the status 'Ready to be recorded'.

### 4.4. Publication

When the lexical entry has been approved for publication by the Signbank team it receives the status 'Ready for publication'. signCollect proceeds to upload the media files and metadata to Global Signbank. The Signbank server reports back whether the upload has been processed successfully. If the upload is successful, the sign is available on the Global Signbank website. The information also remains available within the signCollect system. Any further changes are synchronised between Global Signbank and signCollect.

## 5. Hardware and software components

### 5.1. Hardware

**Sony FX30 Camera** For the recordings, we employ three Sony FX-30 camera units with a standard lens of 36-105mm at an F/4 aperture, a shutter speed of 1/500, and an ISO setting of either 200 or 400. The cameras are positioned four meters away from the signer at angles of 25, 0 and -25 degrees, respectively. They are connected with USB-C to signCollect Studio which facilitates data transfer and operational commands, such as start/stop controls.

**Tentacle Sync E** Each Sony FX30 camera is connected to its own Tentacle Sync device, which generates time-code for synchronization purposes (Tentacle, 2024). The Tentacle Sync devices are all interconnected via Bluetooth, ensuring they remain synchronized with the same time-code during capture, avoiding time drift between recordings of the same sign by different cameras.

**iPhone with Live Link Face app** Besides three video cameras our studio setup optionally includes an iPhone as well, with the Live Link Face app by Unreal Engine (Live Link Face, 2024) which captures facial expressions of the signer and streams the data to signCollect. This data is saved in blendshapes CSV format and is added as metadata of the given sign.

### 5.2. Software components

**Google Mediapipe** The MediaPipe Gesture Recognition component (Mediapipe, 2024) captures and interprets hand gestures in real-time. Hand movements of the signer are being tracked and translated into labels. For instance, if the signer produces a 'Thumb Up' sign, then it is translated as the 'thumbUp' label. For our specific purposes we created a dataset consisting of thumb up, thumb down, and hands down recordings from different angles and different people and trained a Tensorflow Lite Model compatible with the Mediapipe interface. This model successfully recognizes the relevant control gestures.

**Sony Camera Remote SDK** The Sony Camera Remote SDK (Sony, 2024) provides users an ability to control the cameras. It includes example applications and complete documentation. We developed a new application with the SDK as library to give the signer the ability to control (start and stop) the cameras in conjunction with Mediapipe Gesture Recognition. signCollect Studio also executes

checkBattery(), checkISO(), and more functions to regularly check the health settings of the cameras.

**signCollect API** We have developed a signCollect API, which is integrated in signCollect Studio. It allows all components of the signCollect ecosystem to communicate with each other. When the signer proceeds to the next sign, the API collects and displays information about that sign from the database. Before proceeding to record the API checks the health and status of the cameras. Via Mediapipe Gesture Recognition the API checks whether the signer has given a thumbs up or down sign. After the signer has completed a recording the API checks the files and sends them to the file server.

## 6. Issues solved by signCollect

### 6.1. Issue One: Efficient Recording

In our basic set-up, we use three video cameras. Sometimes, we also capture facial data with Live Link Face on an iPhone. Without a system like signCollect this would mean that every recording session would require the presence of at least one person besides the signer to control and manage the devices. Moreover, the settings of all devices such as ISO, resolution, bit values, frame rate, and color values would have to be adjusted and checked manually before each recording session (because the devices are also used for other projects with different settings). After the recording session, all videos would have to be synchronized and edited using editing software. The editor would also have to manually specify the correct Annotation ID gloss for each recorded sign. All of these steps would require a significant time and would be error-sensitive. By automatizing most steps, data collection and quality control can be done much more efficiently.

### 6.2. Issue Two: Collaboration

With the signCollect system, it is easier for the Signbank team to collaborate and coordinate the procedures of selecting, recording, validating, and publishing lexical items. The system provides a fixed procedure to contribute new items to the dataset, avoiding data fragmentation and pollution.

## 7. Possible Future Extensions

We envision that the signCollect ecosystem may be further enriched in various ways in future work. For instance, each sign in Global Signbank has a phonological description, which currently needs to be provided manually. This requires a major time-investment. One possible direction would be to

implement a functionality in the signCollect system for semi-automatic phonological description (see, e.g., [Ranum et al., 2024](#)).

For the identification of new signs to be added, we envision using Gloss Clustering ([Moryossef, 2023](#)) based on the Corpus NGT ([Crasborn et al., 2008](#)), HoReCo ([De Sisto et al., 2023](#)), and possibly other corpora. Gloss Clustering presents a display of unique glosses per video or dataset with frequency counts and information as to whether it is a known or unknown gloss. This methodology can identify signs that could be considered for inclusion in the lexicon. Leveraging metadata from the corpora, the system could also indicate for each identified sign in which years it occurred most and where it was used.

These tools could aid the Signbank team in the identification and documentation of new signs.

## 8. Conclusion

The signCollect system has been developed to enable more efficient and consistent sign language data collection efforts. The main focus of this paper has been on the innovative ‘touchless’ multi-view video recording pipeline, which we have tested extensively. We can record between 60 to 120 lexical entries per hour and publish them within a day on Global Signbank, without the need for technical experts overseeing the recording and post-processing process. This streamlines the work of the Signbank team and offers other teams opportunities to similarly optimize their data collection processes. With feedback from workshop participants, we intend to further develop the system and make it available as an open-source tool for all sign language researchers.

The current codebase can be found at <https://github.com/rem0g/signCollect>

## 9. Acknowledgments

We thank Rob Belleman and Joey van der Kaaij for providing hardware to develop and run the signCollect software, Dalene Venter for assisting in writing the paper and Casper Wubbolts for testing the signCollect Studio. We gratefully acknowledge funding from the Platform Digital Infrastructures for the Social Sciences and the Humanities (PDI-SSH) in the Netherlands.

## 10. Bibliographical References

Steve Cassidy, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen, and Trevor Johnston. 2018. Signbank: Software to support web based dictionaries of sign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Crasborn, Onno and Bank, Richard and Zwitterlood, Inge and van der Kooij, Els and Ormel, Ellen and Ros, Johan and Schüller, Anique and de Meijer, Anne and van Zuilen, Merel and Nauta, Yassine Ellen and van Winsum, Frouke and Vonk, Max. 2020. *NGT dataset in Global Signbank*. Radboud University, Centre for Language Studies, ISLRN 976-021-358-388-6.

Onno Crasborn, Inge Zwitterlood, and Johan Ros. 2008. Corpus NGT. *An open access digital corpus of movies with annotations of Sign Language of the Netherlands*. Centre for Language Studies, Radboud University, Nijmegen.

Mirella De Sisto, Dimitar Shterionov, M DeSisto, D Shterionov, Lien Soetemans, Vincent Vandeghinste, Caro Brosens, and Vlaams Gebarentaalcentrum. 2023. NGT-HoReCo and GoSt-ParC-Sign: Two new Sign Language-Spoken Language parallel corpora. In *CLARIN Annual Conference Proceedings*, page 6.

Thomas Hanke, Reiner Konrad, and Arvid Schwarz. 2001. GlossLexer: A multimedia lexical database for sign language dictionary compilation. *Sign Language & Linguistics*, 4(1-2):171–189.

Thomas Hanke and Jakob Storz. 2008. *iLex – a database tool for integrating sign language corpus linguistics and sign language lexicography*. In *Proceedings of the Third Workshop on the Representation and Processing of Sign Languages at LREC 2008*, pages 64–67.

Ulrika Klomp, Lisa Gierman, Pieter Manders, Ellen Nauta, Gomèr Otterspeer, Ray Pelupessy, Galya

Stern, Dalene Venter, Casper Wubbolts, Marloes Oomen, and Floris Roelofsen. 2024. An Extension of the NGT Dataset in Global Signbank. In *Proceedings of the Eleventh Workshop on the Representation and Processing of Sign Languages at LREC-COLING 2024*.

Live Link Face. 2024. Facial Capture with Live Link | Tutorial — dev.epicgames.com. <https://dev.epicgames.com/community/learning/tutorials/LEYe/unreal-engine-facial-capture-with-live-link>. [Accessed 29-02-2024].

Mediapipe. 2024. Mediapipe - google/mediapipe: Cross-platform, customizable ML solutions for live and streaming media. — github.com. <https://github.com/google/mediapipe>. [Accessed 29-02-2024].

Moryossef. 2023. sign.mt: Effortless real-time sign language translation. <https://sign.mt/>.

Ranum, Otterspeer, Andersen, Belleman, and Roelofsen. 2024. 3D-LEX v1.0 3D Lexicons for American Sign Language and Sign Language of the Netherlands. Submitted for publication. Submitted for publication.

Sony. 2024. Camera Remote SDK | SONY — support.d-imaging.sony.co.jp. <https://support.d-imaging.sony.co.jp/app/sdk/en/index.html>. [Accessed 29-02-2024].

Tentacle. 2024. Tentacle Sync E | tentacle sync — tentaclesync.com. <https://tentaclesync.com/sync-e>. [Accessed 29-02-2024].



# The EASIER Mobile Application and Avatar End-User Evaluation Methodology

Frankie Picron<sup>1</sup>, Davy Van Landuyt<sup>1</sup>, Rehana Omardeen<sup>1</sup>,  
Eleni Efthimiou<sup>2</sup>, Rosalee Wolfe<sup>2</sup>, Stavroula-Evita Fotinea<sup>2</sup>,  
Theodore Goulas<sup>2</sup>, Christian Tismer<sup>3</sup>, Maria Kopf<sup>4</sup>, Thomas Hanke<sup>4</sup>

<sup>1</sup>European Union of the Deaf, Belgium

<sup>2</sup>Institute for Language and Speech Processing, ATHENA Research Center, Greece,

<sup>3</sup>Nuromedia GmbH, Germany

<sup>4</sup>Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Germany

<sup>1</sup>{frankie.picron, davy.van.landuyt, rehana.omardeen}@eud.eu, <sup>2</sup>{eleni\_e, rosalee.wolfe, evita, tgoulas}@athenarc.gr, <sup>3</sup>christian.tismer@nuromedia.com

<sup>4</sup>{maria.kopf, thomas.hanke}@uni-hamburg.de

## Abstract

Here we report on the methodological approach adopted for the end-user evaluation studies carried out during the lifecycle of the EASIER project, focusing on the project's mobile app and avatar technologies. Evaluation was led by deaf consortium partners and performed in two cycles, involving both deaf signers and hearing sign language (SL) experts groups from five SLs to provide user feedback, which served as a reference to base the next development steps of the respective EASIER components. With this goal in mind, priorities were (i) to exploit information gathered via focus group discussions after (ii) presenting evaluators with the technological components and related questionnaires fully accessible to signers to maximize feedback and underline the importance of user involvement in the development of the technology.

**Keywords:** Avatar technology, end-user evaluation, sign language translation mobile application, usability, user acceptance, avatar legibility, sign language accessible questionnaire

## 1. The EASIER Concept

EASIER<sup>1</sup>, a Horizon 2020 project, which ended 31st December 2023, was established with the aim to design, develop, and validate a complete multilingual machine translation system which would act as a framework for barrier-free communication among deaf and hearing individuals, as well as provide a platform to support sign language content creation.

The project concept was based on a unique combination of technological innovations, sign language resources and sign language linguistics expertise, allowing among other for the incorporation of a signing avatar that integrates sign language grammar and prosody features to perform the most advanced synthetic signing currently available, into a mobile application designed to provide users with an easy-to-use translation tool to serve everyday translation needs.<sup>2</sup> Envisioned functionalities of this tool included bi-directional translation between spoken and signed languages (and vice versa), incorporating options for sign, text and speech as both input and output modalities. The EASIER mobile application was tested with five sign-spoken language pairs with the aim to create a flexible framework that could be further expanded to include other languages.

To achieve these goals, user involvement in the development of technologies has been one of the main pillars of the EASIER project. The user-centric approach of the project encompassed continuous involvement of deaf signers and SL experts in the consortium and throughout the project steps. The technology was validated in two end-user evaluation studies, the first one taking place in 2022, shortly after the mid-lifecycle of the project (see Picron, Van Landuyt and Omardeen, 2022) and the final one in 2023, close to the end of the project (see Picron et al., 2023). This paper describes in detail the design and implementation of the final end-user evaluation study of the EASIER project, specifically with respect to the signing avatar and the mobile application technologies. Our focus is documenting the evaluation methodologies in detail, rather than presenting the results, which can be found in Picron et al. (2023).

## 2. The EASIER End-User Evaluation Methodology

The EASIER mobile application and the avatar components were evaluated in a facilitator-led group setting, in sessions which took place both on-line and in situ, where participants were first shown the current state of the technology and asked to complete a structured rating task,

<sup>1</sup> <https://www.project-easier.eu/>

<sup>2</sup> An account of Machine Translation technology developed in EASIER can be found in Müller et al. (2023).

followed by a facilitator-led group discussion to get more in-depth qualitative feedback about the technology. This approach allowed us to not only get global benchmarks for how the technology is viewed by users, but also collect qualitative feedback on how to best improve technologies to achieve maximum user acceptance. For participation of end-users to all evaluation activities, a signed consent form was required, where the consent form content was provided both in text and signing.

## 2.1 Recruitment Strategy

Deaf and hearing participants were recruited from the following sign language communities: British Sign Language (BSL), German Sign Language (DGS), Swiss German Sign Language (DSGS), Greek Sign Language (GSL), and French Sign Language (LSF). For each of the five communities, there were two separate evaluation groups, one with deaf and hard of hearing and one with hearing participants, resulting in a total of 10 groups. Separate deaf and hearing groups were used to create a 'safe space' in which participants could freely and comfortably express themselves in their preferred language among peers. To set up the different focus groups, local project partners identified facilitators and participants for the evaluations. For the deaf groups, a deaf facilitator was chosen and for the hearing group, a hearing facilitator was chosen, while in the case of DGS the facilitator was hard-of-hearing (HoH). For GSL a hearing project member who is a CODA and a long-standing member of the signing community acted as facilitator for the deaf group. For each group, between 5 and 7 participants were recruited who use the target sign language. No specific professional or educational background was required for participants; however, for those evaluating the avatar, a high degree of fluency in the relevant sign language was a requirement. Recruitment was carried out through personal and professional networks, while across all groups, some participants who took part in the interim evaluation were invited back for the final evaluation. This mixture was chosen to have both, the experience of the first round allowing to judge the progress, and "fresh eyes" judging from a neutral perspective. Evaluators' anonymity was preserved since only basic demographic information was shared among technology developers, provided in the form of a cumulative report of findings from all evaluation groups.

## 2.2 Evaluation Setup

Given the scale of the evaluation and the number of partner institutions, each facilitator determined their technical set-up. While most elected to conduct in-person evaluations (see Fig. 1 for a group setup), some groups decided for online (see Fig. 2 for an on-line setup) or mixed evaluations to make recruitment and participation easier. Several partners conducted multiple small

group evaluations to optimize scheduling participants. For those groups that were conducted online, participants used their own devices (either mobile phones or computers) to navigate the online app and app questionnaire as well as the avatar questionnaire. For those evaluations conducted in person, in some cases, participants brought in their own devices and in other cases, they used devices provided by the institutions or a combination of both. In several in-person groups, facilitators also used projectors or large computer screens to provide visuals during the discussion.

For most groups, the evaluation sessions were recorded using either video or audio recording devices. Several groups used wide-angle cameras such as GoPros to record the entire scene. These recordings were then used by facilitators to later compile a report detailing the content of the focus group discussion. Recordings were kept by the local institution and not shared with any other consortium members. In most groups, the facilitator and participants were the only ones present in the room during evaluation, but in some cases technical staff also assisted with video recording of sessions.

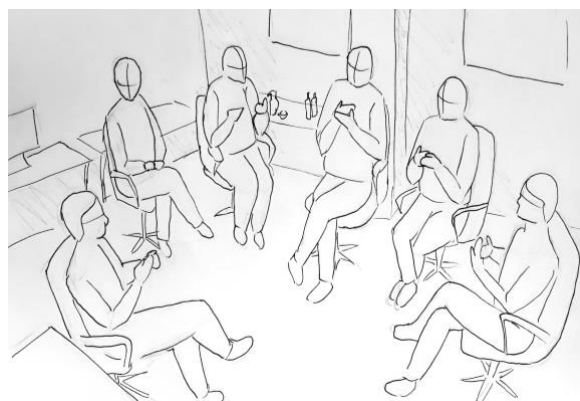


Figure 1: The EASIER group discussion setup.

Furthermore, for some groups, the facilitator for the other group was also present to take notes. Evaluation sessions with deaf groups were conducted in the local sign language, and sessions with hearing groups were conducted in the local spoken language.

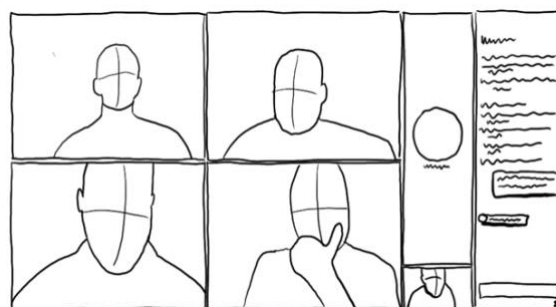


Figure 2: The EASIER on-line evaluation setup.

### 3. The EASIER Mobile Application and Avatar Evaluation

The evaluations consisted of two parts, one evaluating the mobile application, the other evaluating the avatar. Both evaluations consisted of an on-line questionnaire followed by in-depth discussions led by the facilitator.

#### 3.1 The Mobile Application

The design and development of the EASIER mobile application followed a user-centric approach (Abrás et al., 2004; Gulliksen et al., 2003): initial development was based on early feedback received during the user specifications and needs analysis project phase. Continuous feedback from subsequent evaluations and small working group studies with deaf users during the project lifetime guided the development of the mobile application.

The EASIER mobile app is designed to take input of either speech, sign language or text, and translate it into all of these modalities (Fig. 4).

Specific features allow users to personalize the settings for their specific input and output preferences, adjust dark and light modes, and access previous translations in an archive in the app. In the final evaluation, an early version of the mobile application was tested, which incorporated all functionalities that at a later stage supported the app's translation service. The purpose of the evaluation was therefore not to test the quality of the translation system, but instead to get feedback on the design and usability of the mobile application from the target group itself.

The final evaluation took part in three stages. Participants in the evaluation study were first instructed to create an account, and freely explore the application's features. They were then asked to complete an online questionnaire about the application's usability (see Table 1). The questionnaire was based on the System Usability Scale (SUS) (Brooke, 1995), "a reliable, low-cost usability scale that can be used for global assessments of systems usability".

For the purpose of this EASIER evaluation, following Ferreiro Lago et al. (2022), the questionnaire was presented in a bilingual format for both questions and answers (see Fig. 3), with both signed and spoken language for all five language pairs making it fully accessible to deaf evaluators. The group then came together for a discussion which concentrated on major themes regarding the application. These themes were selected based on feedback received in the interim evaluation, and involved the application's (i) settings, (ii) translation, (iii) visual design, (iv) navigation, (v) video recording and (vi) avatar output.

In both evaluation cycles, the mobile application evaluation generated a lot of engaged feedback from end users. The evaluation also added new evidence regarding the ways user preferences and expectations are formulated when

participants are asked to judge the usability of a mock-up or experience the use of a prototype application.

The SUS Questionnaire	
1.	I think that I would like to use this system frequently.
2.	I found the system unnecessarily complex.
3.	I thought the system was easy to use.
4.	I think that I would need the support of a technical person to be able to use this system.
5.	I found the various functions in this system were well integrated.
6.	I thought there was too much inconsistency in this system.
7.	I would imagine that most people would learn to use this system very quickly.
8.	I found the system very awkward to use.
9.	I felt very confident using the system.
10.	I needed to learn a lot of things before I could get going with this system.
Rating options for each question	
-	Strongly disagree
-	Somewhat disagree
-	Neither agree or disagree
-	Somewhat agree
-	Strongly agree

Table 1: The SUS Questionnaire and ratings used in the EASIER application evaluation.

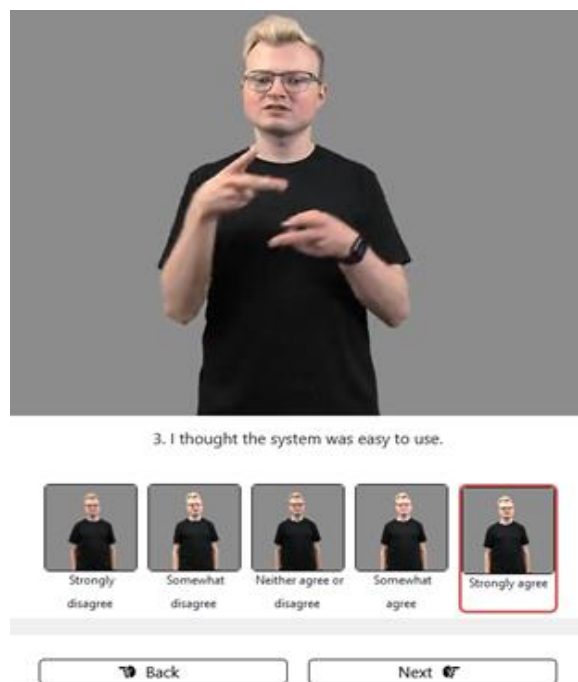


Figure 3: The SUS questionnaire for English and BSL.

While the participants appreciated the variety of input and output options for translation directions, with a mostly straightforward translation process, they demanded some advances. Main points were simplified settings, retranslation feature, searchable archive, enhancements for the video layout (e.g. mirroring, orientation), side by side in-and output (see Picron et al., 2023). These results and a median SUS score of 65 provide a useful benchmark for future work, while the qualitative round of feedback provided useful information on the strengths and weaknesses of the application, providing a roadmap for fine tuning.

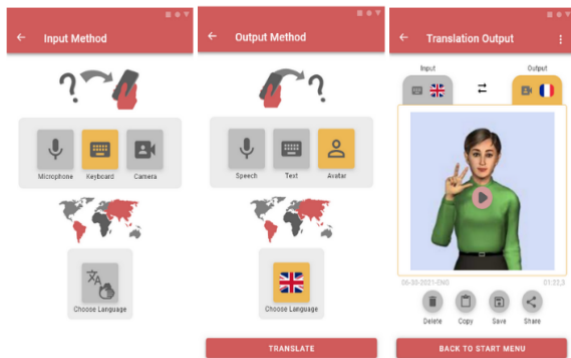


Figure 4: The EASIER app functionalities design for user input/output and translation display.

### 3.2 The Avatar

The main goal of avatar development for the EASIER project was to create fully legible synthetic signing (Wolfe et al., 2022a), with an avatar that was able to incorporate non-manuals, mouthing (Wolfe et al., 2022b), affect, prosody and SL grammar features beyond morphology (Hanke et al., 2023). To test the stages of development and ensure that research work was on the right track, user involvement has been critical. After basing initial work on the user needs analysis conducted in the first project phase continuous ongoing feedback was sought from the signing communities. To reach the avatar users, an on-line multilingual questionnaire<sup>3</sup> was developed, designed to be fully accessible via SL and easily modifiable with respect to content (Dimou et al., 2022b) (see Fig. 5). This questionnaire was used initially in a pilot survey on user preferences, drawing on two well-known avatar engines used in dynamic synthetic signing: the Anna and Paula avatars respectively (Dimou et al., 2022a). It was then adapted for use in the first and second evaluation cycles adding new content for evaluation. Although the questionnaire could be completed anonymously, it also allowed for the option of direct user input via signing into the camera of the user's device (PC/mobile phone), if this was desired.<sup>4</sup>

<sup>3</sup> The current version of the EASIER avatar evaluation questionnaire: <https://sign.ilsp.gr/slt-eval-2/>

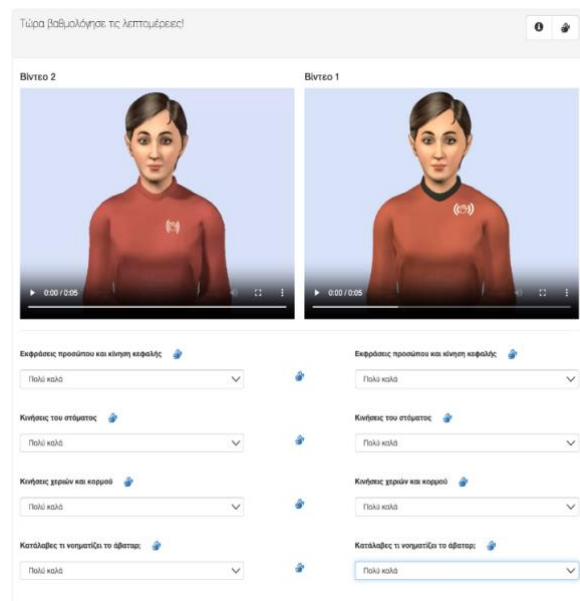


Figure 5: The EASIER avatar evaluation questionnaire (screenshot from GSL/Greek version)

The questionnaire was prepared for the four sign languages for which the avatar was available at the time: GSL, DGS, DSGS and LSF. Thus, eight groups (deaf and hearing from each of the four languages) completed the evaluation procedure. Questionnaires for each language pair were bilingual with both text and sign language and contained signed instructions for navigating each page. Before presentation of the evaluation content, some basic demographic information about the participants was collected, including their age, gender, context of sign language acquisition and self-assessment of their sign language proficiency.

For the final user evaluation, since a major goal was to measure user opinion differentiation with respect to the avatar status during the previous evaluation cycle, participants were presented with a series of screens for each animation. On the first screen, they viewed a video of an utterance produced by a human signer, which was identical in content to the utterance produced by the avatar animations. Then on the next screen, they viewed two avatar animations side by side and were asked a series of questions. Test utterances had the same semantic content across all languages.

First, they had to identify which of the two avatar animations was better. They were then asked to rate the general performance of both animations on a five-point Likert scale ranging from “very good” to “bad”. On the third screen (Fig. 4) they viewed the two animations side by side again and were asked to rate each of them on (1) facial expressions and head movements, (2) mouth

<sup>4</sup> Given that recording of participant video requires special permissions, consent for activation of this specific feature of the questionnaire is also mandatory.



movements, (3) hands and body, and (4) the legibility/intelligibility of the signing. All were rated on a five-point Likert scale, where the options for (1), (2) and (3) were “very good”, “good”, “so-so”, “rather bad” and “bad”, and the options for (4) were “very good”, “good”, “1-2 points were not clear to me”, “it was difficult to understand” and “I did not understand anything” (see Figure 4, text in Greek).

The group discussions following the avatar rating focused on the overall avatar appearance and quality of the animation, the prosody in the signed utterance, the manual signing, the non-manual features and the mouthing of the animation.

In both EASIER evaluation cycles, we used the avatar evaluation questionnaire to assess the legibility and naturalness of the EASIER avatar signing. During the first evaluation cycle, although evaluators were asked to judge the avatar’s hand activity only, they made clear that they wanted to see more facial activity, including mouthing as well as affect. They also wanted to see more prosodic features. These findings prioritized development during the final project period, which was evaluated at the final end-user evaluation cycle. Across all four languages evaluated, user reactions to the avatar’s naturalness and legibility were positive with over 90 percent of user ratings at 3 or above (naturalness rated 3 or above: 92.3%, legibility rated 3 or above: 92.8%).

#### 4. Conclusion

Our findings verified that continuous end-user involvement in SL technology development has proven to be the key for user acceptance and trust of the delivered tools and services. Evaluation cycles which involve larger end-user groups than those involved in a project, provide significant new feedback which is crucial to creating quantitative benchmarks to measure future improvements, while qualitative feedback provides a clear path to improving these technologies in future work. A significant aspect in evaluating SL technology is to provide evaluators with fully SL accessible questionnaires. The feedback received from the EASIER evaluator groups has verified the importance of SL based interfaces and questionnaire content.

Finally, the focus group discussion approach proved to reveal significant aspects of user attitude towards the evaluated technology, also unfolding user expectations and reservations, which the quantitative questionnaire-based approach if adopted as the only method to measure user opinion, cannot bring to light. Thus, the combination of focus group discussion and questionnaire-based evaluation can be suggested as a best practice end-user evaluation method. Finally, it must be mentioned that deaf-led evaluation is a feature that is appreciated by deaf communities.

#### 5. Acknowledgments

This work is fully supported by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union’s Horizon 2020 research and innovation programme, grant agreement n°101016982.

#### 6. References

- Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, (2004) User-Centered Design. In Bainbridge, W. (2004). User-Centered Design. In: Bainbridge, W., Ed., *Encyclopedia of Human-Computer Interaction*, Sage Publications, Thousand Oaks, CA, 445-456.
- John Brooke (1995). [SUS – A quick and dirty usability scale](#).
- Athanasia-Lida Dimou, Vassilis Papavassiliou, Theodoros Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, and Eleni Efthimiou (2022a). [What about synthetic signing? A methodology for signer involvement in the development of avatar technology with generative capacity](#). *Frontiers in Communication*. 7:798644.
- Athanasia-Lida Dimou, Vassilis Papavassiliou, John McDonald, Theodoros Goulas, Kyriaki Vasilaki, Anna Vacalopoulou, Stavroula-Evita Fotinea, Eleni Efthimiou, Rosalee Wolfe (2022b). Signing Avatar Performance Evaluation within the EASIER Project. *7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, 39-44.
- Jan Gulliksen, Bengt Göransson, Inger Boivie, Stefan Blomkvist, Jenny Persson & Åsa Cajander (2003). [Key principles for user-centred systems design](#). Behaviour & Information Technology, Taylor & Francis, Volume 22, 2003 - Issue 6, 397-409 | Published online: 19 May 2010.
- Emilio Ferreiro Lago, María Jesús Pardo Guijarro, and Eva Gutierrez-Sigut, (2022). Diseño de cuestionarios web en investigaciones accesibles para personas sordas mediante herramientas no estándar. *Revista de Estudios de Lengua de Signos*. 4, 29-49.
- Thomas Hanke, Lutz König, Reiner Konrad, Maria Kopf, Marc Schulder, Rosalee Wolfe (2023). [EASIER Notation: A proposal for a gloss-based scripting language for sign language generation based on lexical data](#). *Eighth International Workshop on Sign Language Translation and Avatar Technology*.
- Mathias Müller, Annette Rios, Amit Moryossef, Sarah Ebling (2023). [EASIER final translation system v2](#). EASIER deliverable D4.3.
- Frankie Picron, Davy Van Landuyt, and Rehana Omardeen (2022). [Report on interim evaluation study](#). EASIER deliverable D.1.3.
- Frankie Picron, Davy Van Landuyt, Rehana Omardeen, Eleni Efthimiou, Stavroula-Evita

- Fotinea, Rosalee Wolfe, Theodoros Goulas, Kyriaki Vasilaki, Amit Moryossef, Mathias Müller, Sarah Ebling, and Christian Tismer, (2023). [Report on final evaluation study](#). EASIER deliverable D1.4.
- Rosalee Wolfe, John C. McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Stavroula-Evita Fotinea, and Annelies Braffort (2022a). [Sign Language Avatars: A Question of Representation](#). *Information* 2022, 13(4), 206.
- Rosalee Wolfe, John McDonald, Ronan Johnson, Ben Sturr, Syd Klinghoffer, Anthony Bonzani, Andrew Alexander, Nicole Barnekow (2022b) Supporting Mouthing in Sign Languages: New Innovations and a Proposal for Future Corpus Building. *7th International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, June 2022. 125-129.

# VisuoLab: Building a sign language multilingual, multimodal and multifunctional platform

Christian Rathmann<sup>1</sup>, Ronice Muller de Quadros<sup>2</sup>, Thomas Geißler<sup>1</sup>,  
Christian Peters<sup>1</sup>, Francisco Fernandes<sup>3</sup>, Milene Peixer Loio<sup>3</sup>,  
Diego França<sup>3</sup>

Humboldt-Universität zu Berlin<sup>1</sup>, Universidade Federal de Santa Catarina<sup>2</sup>, Levante Lab<sup>3</sup>  
[christian.rathmann@hu-berlin.de](mailto:christian.rathmann@hu-berlin.de), [ronice.quadros@ufsc.br](mailto:ronice.quadros@ufsc.br), [thomas.geissler@hu-berlin.de](mailto:thomas.geissler@hu-berlin.de),  
[chris.peters@hu-berlin.de](mailto:chris.peters@hu-berlin.de), [francisco.fernandes@levantelab.com.br](mailto:francisco.fernandes@levantelab.com.br), [milene.loio@levantelab.com.br](mailto:milene.loio@levantelab.com.br),  
[dfvdiego@gmail.com](mailto:dfvdiego@gmail.com)

## Abstract

VisuoLab is a multifunctional, multimodal and multilingual platform designed for sign language communities. This platform is based on web accessibility and usability, and is specifically designed in a visual way. All resources are organized to be available in sign languages and written languages for different purposes: to provide materials related to and in sign languages, to produce materials (such as papers, video books, teaching materials that include signing production), to teach with signing tools, to interpret and translate activities for training purposes, and to evaluate signing progress. VisuoLab is designed as an open source platform. The current stage of VisuoLab is a beta version available in the development area of Levante Lab for the platform: <https://visuolab.levantelab.com.br/>

**Keywords:** Visual design, Sign language documentation, Sign language visualization, Sign language learning, teaching and assessment tools, Sign language translation and interpretation tools

## 1. Introduction

The VisuoLab platform<sup>1</sup> was born out of the need to have a robust website built and accessible in sign languages for different purposes. The initial focus was on creating sign language materials, sharing sign language publications, teaching in sign language, and interpreting/translating between sign languages and written and spoken languages. VisuoLab includes interfaces designed for signing communities, following the basic idea of Signbank 2.0. It aims to enable signers with diverse needs by using a visually accessible platform equipped with interfaces based on sign language and to develop a portal and a dashboard in which users can make their own changes, updates, and adaptations in the platform at any time.

The VisuoLab platform is being developed to be an open resource different from some of the previous platforms that we used as a starting point in specific modules.

For the sign language repository, we started from the Portal of Libras (<https://portal-libras.org/>) which was also established based on previous portal improvements considering feedback from its users. The repository is designed to make materials available in sign language, with deaf people as the main target group. It is a place where we bring together publications in sign language. It is an innovative portal in terms of accessibility of materials in sign language and in

having a dashboard that its users can access and manage based on sign language.

Levante Lab also had previous platforms developed for the Brazil Ministry of Education that were implemented to make a more autonomous administration of the tools to be accessible to the users. One drawback of the previous platform is that it requires developers to work on changes to the system that could be done by the administrators of the platform. Working from these previous developments, we applied the two user spaces of the platform for Signbank first (already available for Libras<sup>2</sup> and soon available for other sign languages), and then we improved and implemented the Visuolab, incorporating recent feedback from these platforms and improving the existing prototype. The two spaces, one for the administration of the platform and the other for the final users with different modules, are developed in ways to make the whole platform independent of the developers.

The production of the multimodal materials module has been developed in the platform based on previous experiences of signed video books produced in Brazil, such as Sign Language Acquisition<sup>3</sup>, Libras Grammar<sup>4</sup>, and International Sign Language: Sociolinguistic Aspects<sup>5</sup>. In these previous works, we needed to hire a company to implement the video book. In the context of the platform, we are developing tools for the user to create their own video book or other resources.

<sup>1</sup> Link to the ongoing development of the platform: <https://visuolab.levantelab.com.br/en>

<sup>2</sup> <https://signbank.libras.ufsc.br/en>

<sup>3</sup> <https://libras.ufsc.br/arquivos/vbooks/aquisicao/>

<sup>4</sup> <https://libras.ufsc.br/arquivos/vbooks/gramatica/>

<sup>5</sup> <https://libras.ufsc.br/arquivos/vbooks/internationalsign/>

The development of other Visuolab modules of sign language learning, teaching and assessment and of interpreting and translation was inspired by *ProSign* products (Rathmann et al., 2019), YASLA (<https://web.yasla.de/>), GoReact (<https://get.goreact.com/>) and Moodle-based teaching and learning resources in the Deaf Studies BA and the Sign Language Interpreting MA at Humboldt-Universität zu Berlin (see e.g. Barbeito Rey-Geißler et al., 2018).

Both Yasla and GoReact are commercially available. However, it has been and always will be a challenge for public Higher Education institutions to ensure funding for these products on the regular basis (e.g. in Brazil).

For these modules, we have developed resources of videos with interactions from users, including comments in videos and written form, as well as the use of chats combined with the video, which can be the source of an activity for the class or for translation and interpretation. Student data is not stored on an external server.

The *ProSign* Portal at the European Centre of Modern Languages (ECML)<sup>6</sup> was designed to make available resources for sign language learning, teaching and assessment). These resources follow the European Reference Framework of Teaching Languages<sup>7</sup>.

The resource is a tool for teachers to develop sign assessment activities that give feedback to students on their development in their sign language learning process. Visually accessible ProSign resources integrate sign language education with the Council of Europe's developments in language education within the framework of the Common European Framework of Reference (CEFR). This previous work has also influenced the organization of the materials for sign language teaching purposes on the platform, combined with the previous experience of Moodle-based e-learning resources developed at the Humboldt-Universität zu Berlin in the context of the CEFR (Common European Framework of Reference).

The focus is on promoting autonomous learning and the concept of blended learning. These resources provide students with the opportunity to independently improve their language competencies in DGS by using e-learning-based Moodle tools, regardless of time and place, alongside classroom learning. In the receptive domain, the various e-learning activities include tasks using multiple-choice questions, true/false questions and drag-and-drop questions. In the productive domain, the tasks involve creating a video recording directly in Moodle, followed by a self-assessed test. Furthermore, the integration of

h5p videos in Moodle enables an interactive video learning experience. Moreover, assessments are used for different purposes including self-evaluation and examination of the respective sign language proficiency level, as well as reflecting on one's own sign language competence within the framework of the European Language Portfolio, ELP.

VisuoLab incorporates the expectations of teachers/instructors and researchers working in sign language and interpreting programs regarding the possibilities of creating sign language materials and tools for learning, teaching, assessment and research. These foundations have led to a robust VisuoLab platform, adding the creation and interaction spaces using sign language as the primary language. Some of the types of expectations incorporated are the following: (1) the possibility for the administrator user to add videos in sign language to the platform's menus; (2) the inclusion of markers in sign language videos to indicate the topics of the videos, which makes it easier to know what type of content is explained in a given minute of the video; (3) the recording of videos within the platform to upload research, teaching, assessment and learning content in sign language from pre-defined subcategories, taking into account areas of impact for the deaf communities; (4) the creation of educational materials for the deaf communities within the platform with the video book tool.

The VisuoLab platform then includes a portal and a dashboard. The portal provides a space for the user account that will have specific credentials to access different parts of VisuoLab. It includes (a) general users who can access the materials shared in the portal; (b) teachers who can create a room for each class, organize the class and the assignments for the students, and assess their sign language proficiency; (c) students, who can access the classes, post their assignments, receive feedback, interact with the teacher, and perform self-assessment (including Language Portfolio and Interpreting Portfolio); (d) creators of new resources, who can create a video book, didactic materials, assignments, and other materials for grammars, anthologies, sign language teaching and learning assessments, and interpreting and translation. In the portal, the database is available to everyone. People who have an account can save their materials and access the areas that their profile allows. Permissions are granted by administrators, managers and teachers to students or assistants, and by material creators to co-authors, editors, designers and assistants. The dashboard is designed to build interfaces based on the principle

<sup>6</sup> <https://www.ecml.at/ECML-Programme/Programme2012-2015/ProSign/PRO-Sign-referencelevels/tabid/1844/Default.aspx>

<sup>7</sup> <https://op.europa.eu/en/publication-detail/-/publication/297a33c8-a1f3-11e9-9d01-01aa75ed71a1/language-en>



of autonomy, giving users the power to manage the platform themselves. This dashboard includes a robust set of resources to manage the whole platform in all multifunctional and multimodal interfaces based on sign language.

To accommodate these different purposes, VisuoLab has four main axes: (1) production of sign language content; (2) availability and indexing of sign language content; (3) research, translation and collaborative learning environments. VisuoLab users can produce their own materials in sign language with different tools, creating videobooks, instructional/didactic resources and literary publications, accessing an area for their creation. The VisuoLab platform is being developed using Davidson's (2008) proposal for a technology-mediated collaborative environment. The idea is that participation is based on different sets of theoretical assumptions of knowledge and authority in decentralized systems. According to Wenger et al. (2002), the community of practice uses technological sources to facilitate and amplify the networks of relationships, so that knowledge is learned through creative techniques. The VisuoLab tools are thus easy for users to access, carefully designed to be visual and based on sign language interactions. Following Camargo and Fazani's (2014) proposals, the technical architecture of the VisuoLab platform has been built based on participatory design approaches and its structure was designed with components and interactions that consider the needs of deaf users. Flor (2016) and Fajardo, Parra, and Cañas (2010) highlight the importance of the use of sign language and the use of contextualized visual resources. We also considered Rosenfeld, Morville, and Arango (2015) with respect to information architecture, which includes the design of localized and understandable information environments. The creation of a technologically mediated collaborative environment fits Davidson's (2008) definition of a generation of tools called Humanities 2.0: "Humanities 2.0 differs from the monumental, data-based projects of the first generation not only in its interactivity, but also in its openness to participation, based on a different set of theoretical premises that decenter knowledge and authority" (Davidson, 2008, pp. 711-12).

The technical and pedagogical requirements and the solutions developed for them both required us to work on the development of the proposed technology in partnership with the deaf users. at different stages of the process (from conception, design phase, development and testing). Without this, the specific challenges of the field would not be satisfactorily addressed, given the difficulty of adapting existing development technologies to the specific needs of sign language users.

Technologies bring new forms of learning, an ubiquitous learning. VisuoLab aims to respond to

a pressing need for access to qualified information by educational professionals and the community in general, and to support training processes for deaf and hearing professionals who work with deaf people, from the perspective of disseminating knowledge on this topic. In addition, VisuoLab has dedicated areas for collaboration that allow its users to publish relevant information on the subject, ensuring constant updating and exchange of information and enabling the diversification and articulation of the public service network for the protection and care of the deaf community. In addition, VisuoLab provides tools that promote a training network and/or community of practice, allowing interaction in sign language and writing. The proposal is therefore possible due to the advantages that networks present in "contingents" (Santaella, 2010). According to Santaella (2010), these learning processes occur with the possibility of making VisuoLab a space that allows users to develop this communication skill at any time and place, through different mobile devices. Thus, ubiquity is associated with mobility, which favors pedagogical practices through access to technologies and establishes a new relationship between space and time. Specifically in the context of sign assessment, we follow Geißler and Barbeito Rey-Geißler (2018) and Barbeito Rey-Geißler, Bittner, and Geißler (in. prep.). It took a collaborative approach in which stakeholders and end-users were actively involved throughout the process. The result is a sign language dominated and deaf-friendly platform because sign language users, deaf experts and deaf professionals have been actively involved in the process.

The feedback of users is being collected in a system built for interaction between users and developers along the process. After approving the prototype, the developers work on the implementation, and the users evaluate it using a shared file in which the user can approve or ask for improvements. The final step is to review the changes by the developers for final approval. This system is also associated with videos showing what the users are accessing. Then, they insert their impressions. It is a very efficient tool, and it is systematic throughout the development process. When necessary, we meet with users and developers to clarify the need for improvements. The basic idea has been to make communication between users and developers very efficient because, in previous experience with the development of platforms in Brazil, we learned that this is a key step of the process.

## **2. Content production with focus on sign languages**

The VisuoLab platform has a specific interface for producing sign language materials. This interface arose from the need to publish video-books, teaching and learning materials, signed papers,

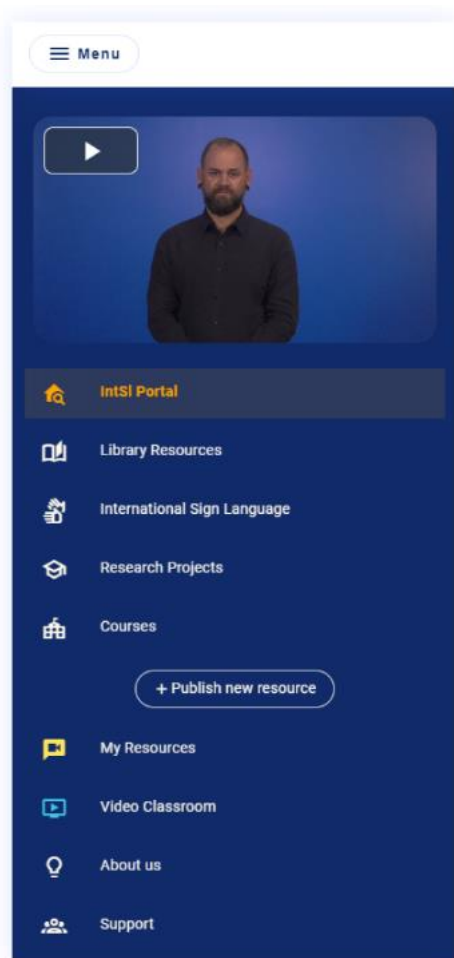


Figure 1: Menu of creation of signed materials

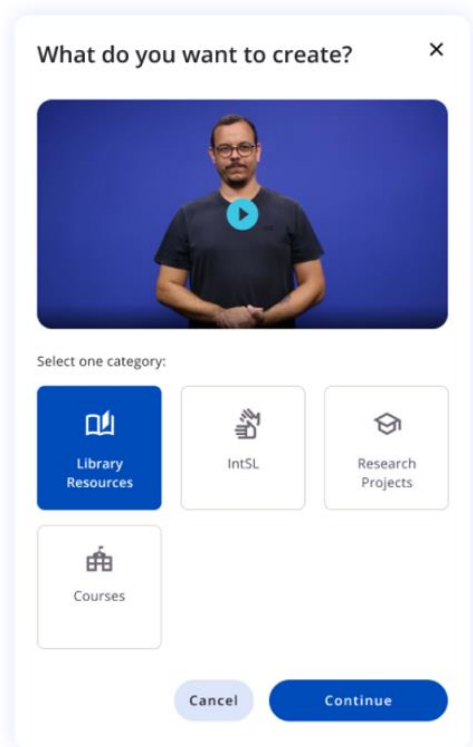


Figure 2: What to create options

and signed videos. The creation of materials allows users to produce work in a multimodal way (sign, written, images, spoken) and to publish their work for the library, in the specific sign language space, as a research project or a course. Each of these categories is organized with tools to support creation with simple interfaces. Figure 1 shows the creation menu and Figure 2 shows the creation options.

The creation interfaces enable users to develop their sign language materials. This space is a collaborative space where authors and researchers can build their publications using sign language resources in addition to the written form. It is also possible to include more than one sign language if translations are available. These materials can be kept as drafts until they are saved. The saved materials can be published on the platform or in other places. They can also be used as resources in the course for teaching and learning within the framework of the collaborative interface.

### 3. Accessing the portal content

The portal includes a library resources hub which covers all the sign language related materials. It is also a site for grammar, anthologies, teaching, learning and assessment sign language, and when available, it may include Signbank sign language corpora, and glossaries of a particular

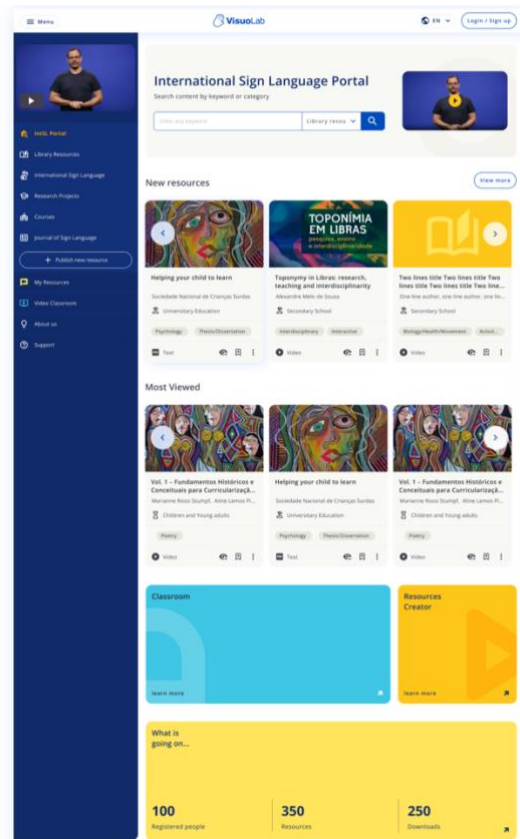


Figure 3: General overview of the platform

sign language. Figure 3 shows the view of the International Sign Language Platform as an illustrative example.

Users can search by selecting the parts of the platform, and they can have a general view of the latest materials and the activity of the different axes of the platform (Figure 4).

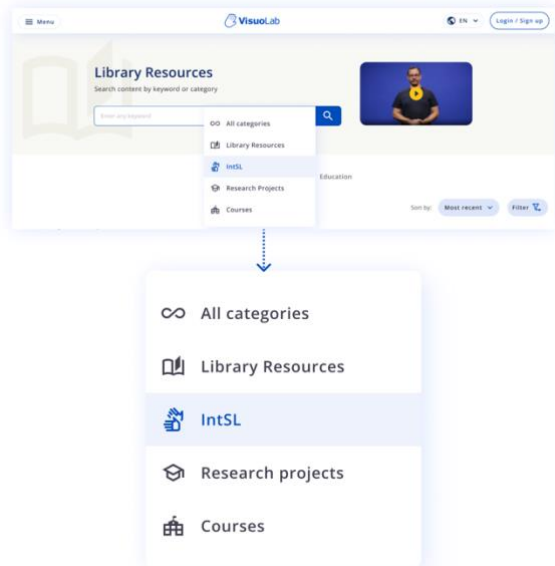


Figure 4: Searching options of the platform

The search can be done by entering a keyword or by category. It is also possible to view the materials by choosing to view all, or by selecting the categories: *Library*, *IntSL* (or the local sign language of the platform), *Research Projects* and *Courses*. It is possible to sort by most recent, most used and saved.

The login at the top right gives the list of spaces that the specific user has. Administrators and managers have access to the dashboard. The menu shows the general parts of the platform and

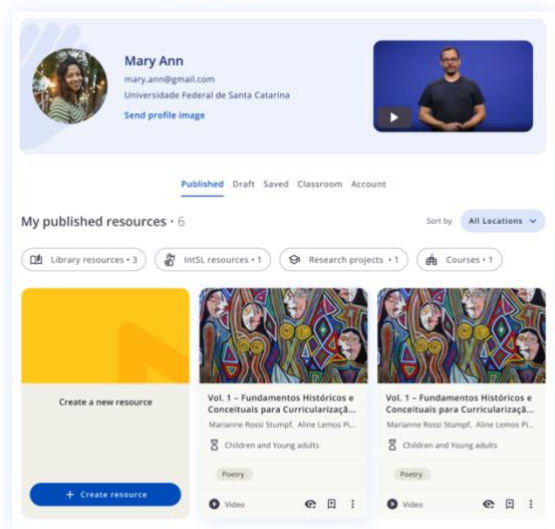


Figure 5: My resources area

the resources of the specific users, where they can directly access all the things they are working on in the platform (for example, classes, interpreting and translation materials, published materials, unpublished or saved drafts, etc.). Figure 5 shows this space.

In the following section, we present the other two spaces that include research, interpreting & translation and collaborative learning environments (courses, creation of resources, and the video classroom).

#### 4. Research, Interpreting & Translation and Collaborative Learning Environments

The VisuoLab platform has multifunctional and multimodal applications that include interfaces for research, collaborative learning, teaching and assessment environments, and translation and interpreting practices. The collaborative learning environments include classrooms with assignments designed for teaching and learning sign language. Teachers can create spaces for classroom activities in sign language, and students can access them and post their answers in sign language. Teachers or students can then add comments, feedback, and suggestions in any sign language activity including formal assessment tasks within the video itself. All these insertions can be accessed directly in the video or in the list of comments related to the activity. The access to these comments is made through an interface that includes markers directly on the video or through the list of comments that may be available in videos or written messages, as shown in figure 6.

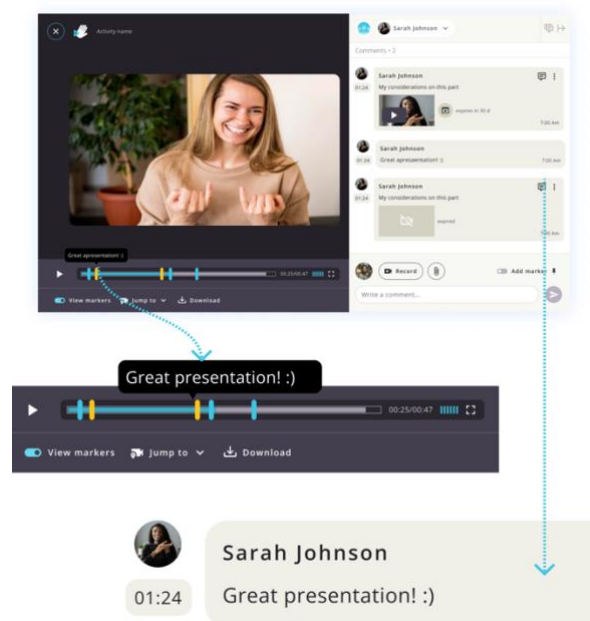


Figure 6: Interface of the learning and translation collaborative work

This interactive collaboration makes it possible to improve the videos and produce new versions when necessary and can also be used for interpreting/translation purposes. The tool includes viewing the video text (signed, written or spoken) in the source language and interpreting/translating this into the target language (sign, spoken or written languages). At the same time, a new video in the target language will be recorded. The interpreted/translated products can also receive comments, suggestions and specific feedback related to different parts of the work done in the video itself. As with the signing activities, there is the option to produce new videos to improve the students' interpreting and translation skills.

In addition, teachers can access the assessment area, including tools of reception, production, and interaction/mediation. Students can use specific tools for self-assessment of their language proficiency in reception, production, and interaction/mediation. Figure 7 shows this interface.

The Video Classroom interface includes rooms (classes), activities and resources. Teachers can see their individual classes and have an overview of what is going on in all their class with the total

number of rooms, the total number of activities and the total number of comments. In this interface, the teacher creates new rooms and activities and adds resources.

Room creation includes video lessons, activities, and resources. When the room is created, the teacher adds the participants as students. Students access their lessons through the class(es) created by the teacher, where they can access the activities and post their assignments in sign language or in writing, as needed. The teacher can provide feedback directly to the student in the video in sign language or by adding written notes.

In the creation of library resources, the user has a choice of three categories: *Literature*, *Academic publications* and *Instructional materials*. The sign language resources allow the user to create materials for the corpus, sign bank, grammar, sign language teaching and learning as well as sign language glossaries. These materials can be used to create courses designed at different levels of instruction. The last area of creation is *Research Projects*, where researchers can prepare their publications, including signed papers and signed examples of their publications.

## 5. Final Considerations

The VisuoLab platform is an open-source platform designed in a community of practice. It is currently a prototype, at a testing stage. Deaf teachers, experts and professionals have been involved in the development of VisuoLab. The result of this process aims to be a sign-language friendly and deaf-friendly platform that includes visual interfaces based on sign language communication. It is a multimodal platform with multifunctional and multimodal resources that make it a robust and complex system. It is designed for the creation of content-based materials on sign languages, interpreting and translation as well as on research collaborative learning environments and sign language assessment. The collaborative platform is the underlying concept used to synthesize the interfaces available.

VisuoLab aims to be a robust platform with multifunctional modules inspired by different sources, but all in one place. The leaders of the concept of the platform are mainly deaf professionals working in sign language studies (sign linguistics, sign language learning/teaching/assessment, sign language translation and interpretation): Christian Rathmann, Thomas Geißler and Chris Peters from Humboldt-Universität zu Berlin, Peter Romanek from Tallinn University/Humboldt-Universität zu Berlin and Ronice de Quadros from Universidade Federal de Santa Catarina. Other deaf professionals and students are accessing

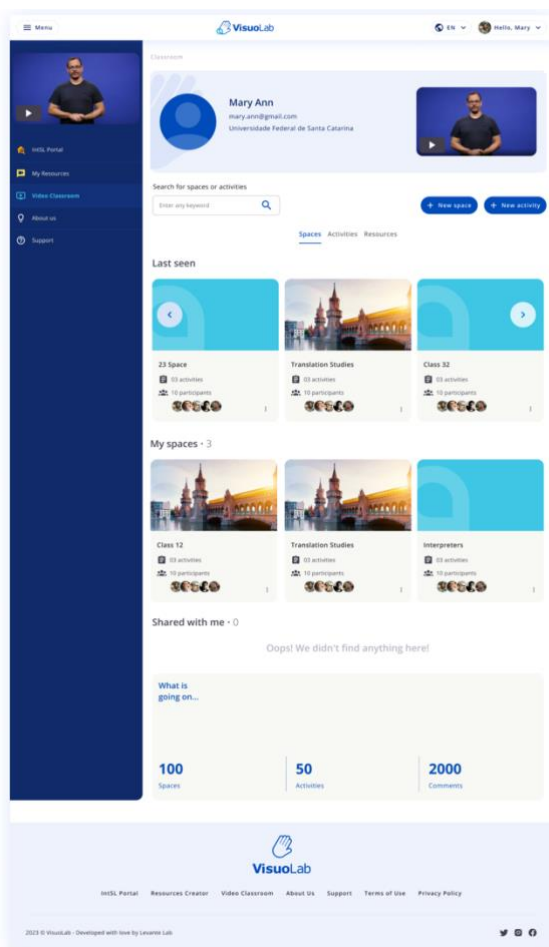


Figure 7: Interface of the classroom



the prototypes for usability purposes during the development process: (1) A quali-quantitative questionnaire was sent to deaf and hearing users to identify the real needs of this audience in relation to a learning platform; (2) Benchmarked research was raised for all the important points for the deaf users that other platforms in the area offer to support the development of Visuolab.

The platform is in full development of its beta version at the moment. Some of the modules and features are available and are being tested with deaf users.

In summary, the development methodology implemented follows the pattern of approaching the deaf users from the beginning of the technology design process. Understanding the needs of users, the niche and the public is the first step in this process. The creation step included the definition of the strategies including all the necessary parts to carry out the project. Our focus is on designing the ideal solution, making use of collaborative creation and evaluation tools that help during this process. Then we will apply the Style Guide, Site Map, Prototyping (low, medium, and high fidelity), and Usability Tests. With the tests carried out, the development of the scripts begins. The prototype is mature enough to be implemented, allowing programmers to code and give materiality to the project.

Along these steps, we implemented the procedural development of a platform with a view to overcoming the challenge of combining pedagogical and technical objectives with epistemological respect for the reference area and with a focus on helping the deaf users in their specific learning processes.

Other platforms for this same audience were developed, achieving good product results for the Brazilian education ministries.

Regarding linguistic experiences, the platform is designed with multimodal interfaces. The administrator and the user spaces may include videos in sign language and written text explaining the functions of each tool. This facilitates the understanding of the tools available. Also, videos in sign language can be added in all modules, making the platform a signing environment.

Finally, there are still technical limitations that we are working on. The main limits are related to the limited availability of conversion libraries and the massive storage and distribution of videos on a large scale. Large companies, such as Vimeo, YouTube, and others, have a certain monopoly on these tools and charge for their use. This problem will be overcome through the development of a mass video conversion service with specific features for technology projects aimed at deaf people.

Interested readers can access the developing area of Levante Lab for the platform: <https://visuolab.levantelab.com.br/> This area is in development, and it is subject to instability, which is why we have not included it in this short paper that has the goal to introduce the novelty of the VisuoLab.

## 6. Acknowledgements

The current research was funded by Department of Deaf Studies and Sign Language Interpreting, Humboldt - Universität zu Berlin, and supported by the National Council for Scientific and Technological Development - CNPq (# 303096/2022-5). Thank you to Rachel Sutton-Spence and the anonymous reviewers for relevant comments for this final version.

## 7. Bibliographical References

- Barbeito Rey-Geißler, P., Bittner, A., and Geißler, T. (in prep.). *Sign Language Production - Self Assessment for Sign Language Learners*.
- Camargo, L. and Fazani, A. (2014). Exploring the Participatory Design as a Support During the Development of Information Systems. In *JInCID: R. Ci. Inf. e Doc.*, Ribeirão Preto, 5(1): 138–150.
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg, available at [www.coe.int/lang-cefr](http://www.coe.int/lang-cefr)
- Davidson, C. (2008). Humanities 2.0: Promises, Perils, Predictions. In *PMLA* 123(3):707–17.
- Fajardo, I., Parra, E., Cañas, J. J. (2010). Do sign language videos improve web navigation for deaf signer users? In *Journal of Deaf Studies and Deaf Education*, 15(3):242–262.
- Flor, C. da S. (2016). *Recomendações para a criação de pistas proximais de navegação em websites voltadas para surdos pré-linguísticos*. 2016. 336 f. Tese (Doutorado) – Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis.
- Geißler, T. and Barbeito Rey-Geißler, P. (2018). E-Learning in der universitären Gebärdensprachlehre. *Das Zeichen*, 109:252–265.
- Quadros, R. M. de. and Krusser, D. and Saito, D. (2022). [Libras Portal: A Way of Documentation, a Way of Sharing](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 48–52, Marseille, France. European Language Resources Association.
- Rosenfeld, L., Morville, P., Arango, J. (2015). *Information architecture: for the web and beyond*. O'Reilly Media, 4th edition.
- Rathmann, C. and ProSign-Team. 2019. *Excellence in Sign Language Instruction*. European Centre of Modern Languages.

Council of Europe, available at <https://www.ecml.at/Thematicareas/SignedLanguages/ProSign/tabid/4273/language/en-GB/Default.aspx>

Santaella, L. (2010). *A ecologia pluralista da comunicação: conectividade, mobilidade, ubiquidade*. São Paulo: Paulus.

Wenger, E., McDermont, R. and Snyder, W. M. (2002). *Cultivating Communities of Practice: a guide to managing knowledge*. Boston, Massachusetts: Harvard Business School Press.

# 3D-LEX v1.0

## 3D Lexicons for American Sign Language and Sign Language of the Netherlands

O. Ranum<sup>1</sup> , G. Otterspeer<sup>1</sup> , J.I. Andersen<sup>1</sup> ,  
R.G. Belleman<sup>2</sup> , F. Roelofsen<sup>1</sup> 

University of Amsterdam

<sup>1</sup>: Institute for Logic, Language and Computation, University of Amsterdam

<sup>2</sup>: Computational Science Lab, Informatics Institute, University of Amsterdam

oline.ranum@student.uva.nl, {g.otterspeer, j.andersen, r.g.belleman, f.roelofsen}@uva.nl

### Abstract

In this work, we present an efficient approach for capturing sign language in 3D, introduce the 3D-LEX v1.0 dataset, and detail a method for semi-automatic annotation of phonetic properties. Our procedure integrates three motion capture techniques encompassing high-resolution 3D poses, 3D handshapes, and depth-aware facial features, and attains an average sampling rate of one sign every 10 seconds. This includes the time for presenting a sign example, performing and recording the sign, and archiving the capture. The 3D-LEX dataset includes 1,000 signs from American Sign Language and an additional 1,000 signs from the Sign Language of the Netherlands. We showcase the dataset utility by presenting a simple method for generating handshape annotations directly from 3D-LEX. We produce handshape labels for 1,000 signs from American Sign Language and evaluate the labels in a sign recognition task. The labels enhance gloss recognition accuracy by 5% over using no handshape annotations, and by 1% over expert annotations. Our motion capture data supports in-depth analysis of sign features and facilitates the generation of 2D projections from any viewpoint. The 3D-LEX collection has been aligned with existing sign language benchmarks and linguistic resources, to support studies in 3D-aware sign language processing.

**Keywords:** Sign Language, Computer Vision, Datasets

## 1. Introduction

Sign language processing (SLP) is a dynamic research area concerned with advancing computational methods for sign languages (SL). This multidisciplinary field encompasses tasks such as the automatic understanding, recognition, translation and production of sign language, contributing to a more inclusive future in language technology.

Despite receiving increased attention across computer sciences (Koller, 2020; Rastgoo et al., 2021), SLP remains less developed compared to other areas within Natural Language Processing (Yin et al., 2021). A significant factor contributing to this disparity is the lack of large-scale, high-quality, and publicly accessible sign language corpora (Bragg et al., 2019). Notably, the majority of these datasets are recorded with cameras that view signers from a single, (near-)frontal perspective (Ali et al., 2022). This scarcity of data impedes modern machine-learning algorithms from learning robust sign representations grounded in the three-dimensional nature of sign languages.

Literature supports that depth-awareness and viewing angle matters in both human (Watkins et al., 2024) and machine (Gao et al., 2023; Rastgoo et al., 2020) SL understanding. This implies that repre-

sentations should reflect a degree of 3D awareness, or risk reduced accuracy under normal real-world conditions, such as non-frontal viewpoints.

While systems such as OpenPose (Cao et al., 2021) enable the estimation of 3D poses from video

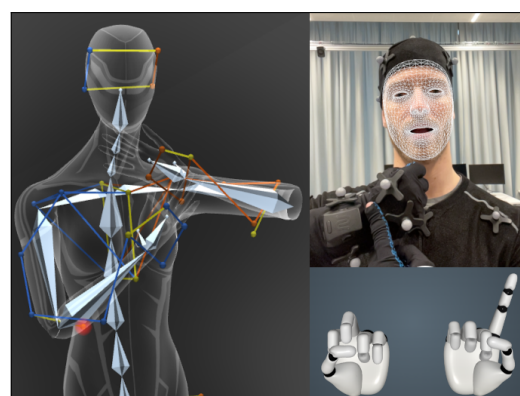


Figure 1: **Motion capture techniques:** The NGT sign *'mango'* captured with the three collection techniques. Left: Pose data captured with Vicon Motion Capture displayed in Shogun Live; Top right: face features captured with Live Link Face (Epic Games); Bottom right: handshapes captured with gloves displayed in Hand Engine (StretchSense).

footage, the precision of such reconstructions is in principle lower than the accuracy achieved through direct 3D motion capture techniques (Jedlička et al., 2020). Navigating imperfectly reconstructed 3D representations can pose significant challenges for downstream SLP tasks.

Providing a 3D ground truth to existing datasets could significantly improve the feasibility of many SLP tasks. Against this backdrop, we introduce 3D Lexicons (3D-LEX) for American Sign Language (ASL) and the Sign Language of the Netherlands (Nederlandse Gebarentaal, NGT). The 3D-LEX datasets include 1,000 isolated signs from each language recorded with three distinct motion capture techniques, as illustrated in Figure 1. The vocabularies have been aligned with existing SL resources, including the WLASL (Li et al., 2020) and SEMLEX (Kezar et al., 2023) benchmarks for isolated sign recognition, the ASL-LEX 2.0 (Sehyr et al., 2021) lexicon and the SignBank NGT (SB NGT) lexicon (Crasborn et al., 2020). The 3D-LEX dataset facilitates the generation of 2D projections from any viewpoint and supports in-depth analysis of sign language features, offering several key advantages:

**Automatic recognition of phonetic properties:** High-resolution 3D data allows for detailed studies of sign language features, including handshapes, place of articulation, and orientations.

**Multi-view SL recognition:** Ground truth 3D representations facilitate the rendering of synthetic multi-view 2D data from any angle and translation. This data can be used to train models that are capable of *multi-view SL recognition*, a task that has received little attention in the SLP literature so far.

**SL production for XR applications:** Current work on SL production focusing on 2D outputs, such as synthetic photorealistic videos or 2D skeleton animations, are not directly suitable for Extended Reality (XR) applications. While reconstructing 3D motion from multiple 2D views is an area of active research, leveraging 3D data to produce 3D animations currently still offers a more effective and accurate approach.

The 3D-LEX v1.0 dataset was developed during our initial exploration of motion capture equipment for capturing three-dimensional sign representations. We acknowledge that the methodology outlined in Section 3 presents significant opportunities for improvement. Specifically, ensuring consistency in data quality will be a primary objective in our future efforts. Nevertheless, even in this nascent stage of development, we could demonstrate the utility of the 3D-LEX data. In Section 4 we showcase how the dataset can be leveraged to produce semi-automatic annotations of handshapes. Evaluating the annotations in a downstream isolated sign

recognition (ISR) task demonstrates that the labels achieved parallel benefits to leveraging annotations provided by linguists. We discuss several observed limitations and prospects for improvement in Section 5, and Section 6 highlights some ethical considerations.

## 2. Background

### 2.1. Sign Language

Sign languages are visual, complete, and natural languages, each with a distinct structure, grammar, and lexicon. They employ a combination of manual markers (e.g. handshapes, hand location, palm orientation and movements) and non-manual markers (e.g. mouthings, facial expressions, gaze) to convey meaning (Stokoe). Sign languages serve as the primary language in Deaf communities.

### 2.2. Sign Language Datasets

The majority of publicly available resources demonstrating sign language are captured in video. These datasets consist of either isolated signs (e.g. Sehyr et al., 2021; Athitsos et al., 2008; Kezar et al., 2023; Joze and Koller, 2019; Li et al., 2020) or continuous sign sentences (e.g. von Agris and Kraiss, 2010; Schembri et al., 2013). Key distinguishing features between the collections include the source language, signer variability, data scope, linguistic domain, and the availability and quality of annotations.

Most datasets comprise RGB video formats, but they may also include depth estimations or skeletal poses generated from joint approximations. While these datasets usually feature a single, (near-)frontal viewpoint, there is a growing trend in lab-curated datasets to provide multiple viewing angles (e.g. Duarte et al., 2020; Mopidevi et al., 2023; Rastgoo et al., 2020; Gao et al., 2023). Depth cameras have been used to capture 3D positioning, for example using the Kinect depth sensor (e.g. Oszust and Wysocki, 2013; Cooper et al., 2012; Huang et al., 2018). For an extensive summary of sign language datasets, refer to Kopf et al. (2022).

Datasets facilitating 3D awareness in sign representations either leverage depth estimations or 3D reconstruction techniques. For the creation of more precise 3D representations, numerous motion capture datasets have been curated (e.g. Lu and Huenerfauth, 2010; Heloir et al., 2006; Benchiheub et al., 2016), typically to generate signing avatars (Bragg et al., 2019) or for exploring automatic synthesis of sign language utterances (e.g. Jedlička et al., 2020; Gibet, 2018).



### 3. The 3D-LEX Dataset

#### 3.1. Data Scope

The 3D-LEX v1.0 dataset includes lexical datasets sampled from ASL and NGT, where the scope was defined to ensure integration with existing benchmarks. A total of 1,000 signs are selected from each language, and recorded with two data collection techniques to capture manual markers and one technique to capture non-manual markers<sup>1</sup>. We release three distinct data formats corresponding to the different capturing techniques, and one component integrating handshapes and body pose data.

**Handshape Data** The handshape(s) of each sign is captured with the StretchSense Pro Fidelity Motion Capture Gloves<sup>2</sup>. The gloves measure the splay and bend of the fingers, alongside the relative rotation of each joint within the hand. The available data include the stretch sensor readings and exported FBX<sup>3</sup> files. Detailed guidance on interpreting and assessing StretchSense data can be found in the project’s Git repository for data evaluation<sup>4</sup>.

**Body Pose Data** The place of articulation, movement, and body pose of each sign is captured using a Vicon (V) Motion Capture setup with optical markers. The raw marker location data is published, alongside processed FBX data, which has been exported via Shogun Post.

**Face Blendshape Data** Facial features are captured as blendshapes with the Live Link Face<sup>5</sup> (LLF) application and ARKit on iPhone.

**Retargeted Animation Data** For sign language production and animation purposes, we release FBX files containing the body pose data and the handshapes.

#### 3.2. Production Method

To efficiently capture the lexicons, we have developed a recording pipeline that achieves an average capture time of 10 seconds per sign. This includes the time for sign demonstration, performance, recording, and storage of the captured sign,

though it varies with the sign’s length. Setup preparations, which involve fitting the suit, positioning markers, and calibrations, require approximately 1 hour with our current method.

##### 3.2.1. Recording Setup and Procedure

Our studio setup includes a designated detection zone for the Vicon cameras, an iPhone equipped with Live Link Face mounted on a tripod, one screen to display glosses and reference videos, and a second screen to showcase the recordings for immediate evaluation.

A triple-foot pedal system facilitates the remote operation of the motion capture control system. Each pedal is configured for a distinct function: The left pedal triggers the start and stop of recordings across all three motion capture systems simultaneously; the middle pedal stores the latest recording and issues a request to the SignCollect platform to display the next gloss in the vocabulary; and the right pedal is used to proceed to the next sign without saving any data. Signcollect is a platform developed to enable the efficient processing of glosses, providing a studio interface managed by gesture or pedal control. For details on the SignCollect platform consult [Otterspeer et al. \(2024\)](#).

The capture process for a single sign involves the following steps: First, the signer assumes an upright posture, with arms relaxed at their sides in a neutral position. By pressing the right pedal, a sign is prompted from the SignCollect platform, and the sign’s gloss and a reference video are displayed on one of the screens. A recording is started by pressing the left pedal, and the signer performs the sign and returns to the neutral stance before the recording is ended with another press of the left pedal. The recorded data is automatically exported to SignCollect and visualized on an avatar rendered with Unreal Engine v5.3, allowing the signer to immediately review the quality of the data. If the data’s quality is satisfactory, the signer can advance to the next gloss by pressing the right pedal, which saves the preceding recording. Should the sign’s execution be deemed inadequate, the signer can repeat the recording by pressing the left pedal again or proceed by pressing the right pedal. For visualizing the sign we created an avatar in *Ready Player Me Studio*, a cross-platform avatar generator that allows users to build avatars for general purposes.

A total of five signers contributed to capturing the ASL and NGT vocabularies. The signers were given two options to operate the pedal. Either they could control the pedal and capture process themselves, or they could delegate the pedal control to a team member. Preferences varied, with three signers opting for controlling the pedal themselves and two preferring assistance to concentrate on signing. Details regarding the number of words

<sup>1</sup>The data is available under a [CC BY 4.0 license](#) at [osf.io/g7u9c/?view\\_only=8090319e12aa4fd991d81e369a1cbd88](https://osf.io/g7u9c/?view_only=8090319e12aa4fd991d81e369a1cbd88)

<sup>2</sup>[stretchsense.com/mocap-pro-fidelity-glove-2/](https://stretchsense.com/mocap-pro-fidelity-glove-2/)

<sup>3</sup>A 3D model file facilitating the transfer of animation data between various modeling applications including Maya, Blender, and Unreal Engine.

<sup>4</sup>[github.com/OlineRanum/SAPA](https://github.com/OlineRanum/SAPA)

<sup>5</sup>[apps.apple.com/us/app/live-link-face/id1495370836](https://apps.apple.com/us/app/live-link-face/id1495370836)

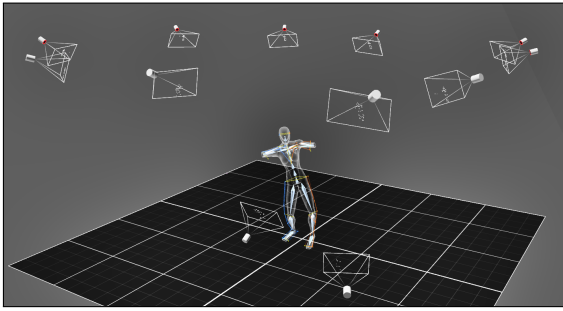


Figure 2: **Setup of the Vicon detection zone:** The illustration indicates the placement of the Vero Cameras on the rig and in front of the signer.

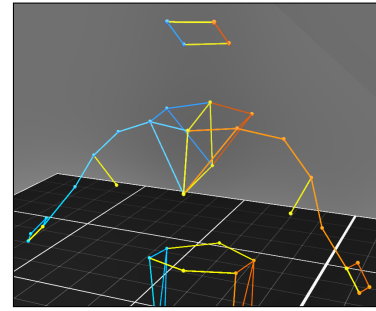


Figure 3: **Marker layout for the Vicon system:** Layout according to FrontWaist 53-marker set template, displayed on signer in Shogun Live.

recorded by each signer per language and pedal control preferences are provided in table 1.

The control system and comprehensive details about the pipeline are available on GitHub<sup>6</sup>. In the following paragraphs we describe each motion capture component in greater detail.

### 3.2.2. Vicon Motion Capture System

*Setup:* A Vicon rig is affixed to the ceiling, equipped with ten Vicon Vero v2.2 optical motion capture cameras<sup>7</sup>, as detailed in Figure 2. To mitigate occlusions, particularly those caused by the lower hands of the signer, an additional two Vicon Vero cameras are placed on the floor in front of the signer.

The markers are placed on the signer following the standard Vicon FrontWaist 53-marker set template<sup>8</sup>, as displayed in Figure 3. Shogun Post is used to make a retarget for the motion capture data, which is used during recording to stream the data to Unreal Engine from Shogun Live.

*Calibration:* For calibrating the Vicon camera system, we adhere to the built-in calibration protocol provided by Vicon. To ensure consistency in the calibration and that the origin remains approximately in the same position across multiple recording sessions, we place masking tape on the floor. This tape serves a dual purpose: one set of markings indicates the precise location for positioning the calibration wand during each calibration process. Another set of tape strips marks the designated spot where the signer is to stand during recordings.

*Software Specifications:* To manage the Vicon camera system, we utilize Shogun Live 1.11, and to perform the retarget of the motion capture data we use Shogun Post 1.11.

<sup>6</sup>[github.com/OlineRanum/GLEX\\_Controller](https://github.com/OlineRanum/GLEX_Controller)

<sup>7</sup>[vicon.com/hardware/cameras/vero/](https://vicon.com/hardware/cameras/vero/)

<sup>8</sup>[docs.vicon.com/display/Shogun18/Create+subjects#Createsubjects-PlaceMarkersPlacemarkersonaperformer](https://docs.vicon.com/display/Shogun18/Create+subjects#Createsubjects-PlaceMarkersPlacemarkersonaperformer)

### 3.2.3. StretchSense Gloves

*Setup:* The StretchSense Pro Fidelity gloves interface with Hand Engine Pro through two USB dongles, which are docked on a separate Dell Universal Dock (UD22) to ensure adequate power supply. Hand Engine is configured to receive remote triggering from Shogun Live, and to retarget animation data directly to Unreal Engine.

*Calibration:* The StretchSense Pro Fidelity gloves are calibrated using the calibration functionality of the Hand Engine software, which involves capturing pre-defined hand poses, to match the recorded output to an individual's hand. Our procedure combines general-purpose poses with specialized ones to customize the glove's fit for each user to capture sign language.

- i. **Express Calibration Poses:** Our general-purpose hand pose set corresponds to the express calibration poses available in Hand Engine, which comprises five common hand-shapes.
- ii. **Advanced Calibration Poses:** A more detailed hand pose library was developed, incorporating the most commonly occurring hand-shapes found in the 3D-LEX NGT (20 poses) and ASL (25 poses) vocabulary, as labeled by linguists in the aligned resources. The advanced pose libraries have been made accessible on GitHub.

We employ the training functionality of Hand Engine to fit the gloves' output data specifically to the signer. We configure all calibration poses to the blend pose mode, a Hand Engine feature that uses the calibration poses as landmarks in a continuous motion space, and interpolates between these poses to yield continuous outputs. The gloves are calibrated and retrained each time a signer puts them on to maintain accuracy.

Following initial consultations with StretchSense about employing the Pro Fidelity gloves for sign

language capture, we developed the specific number of poses and this calibration scheme. However, throughout the creation of 3D-LEX and subsequent discussions, it became evident that the calibration scheme was not ideal. We acknowledge this shortcoming and will reevaluate the calibration process in future works. For a discussion of these limitations and suggestions for potential improvements, please see Section 5.

**Software Specifications:** The StretchSense Pro Fidelity gloves are operated with the Hand Engine Pro software, version 3.0.6.

### 3.2.4. Live Link Face

**Setup:** An iPhone is mounted on a tripod, which is placed directly in front of the signer. Recordings are started, stopped, and saved automatically by the remote triggers.

**Calibration:** Live Link Face was not calibrated per signer. However, this functionality is available in the Live Link Face application and should be explored in a later version of the dataset.

**Hard- and software Specifications:** We use an iPhone 13 Pro and run Live Link Face version 1.3.2 with iPhone AR Kit.

### 3.3. Dataset Characteristics

The recording procedure introduces several recurring patterns into the raw data. Notably, the initial and final arm and hand positions often adopt a neutral stance, with the handshape closely resembling a ‘5’ handshape (refer to Figure 7). This results in, for instance, parts of the handshape recordings capturing signals that are not characteristic of a particular sign. This includes handshapes observed during the transition from a neutral state to the sign’s active posture, or when a sign involves a series of distinct handshapes, resulting in recordings that capture multiple pose signals within a single sign. An illustration of a typical temporal series according to the Euclidean distance is provided in Figure 5.

Data captured using LLF presents a non-uniform sampling rate, as frames are only recorded upon detected changes in the current state of the sensor. Conversely, the body poses captured with the Vicon system and handshapes captured with StretchSense are sampled uniformly.

The lexicons include a variety of handshapes. Figure 4.a showcases the distribution of handshapes in the ASL Lexicon, annotated by sign language linguists in the ASL-LEX resource.

**Signer characteristics** All participants are native signers, who acquired sign language from an early age. Details about each signer’s primary language,

along with their preferences for operating the pedal, are provided in Table 1.

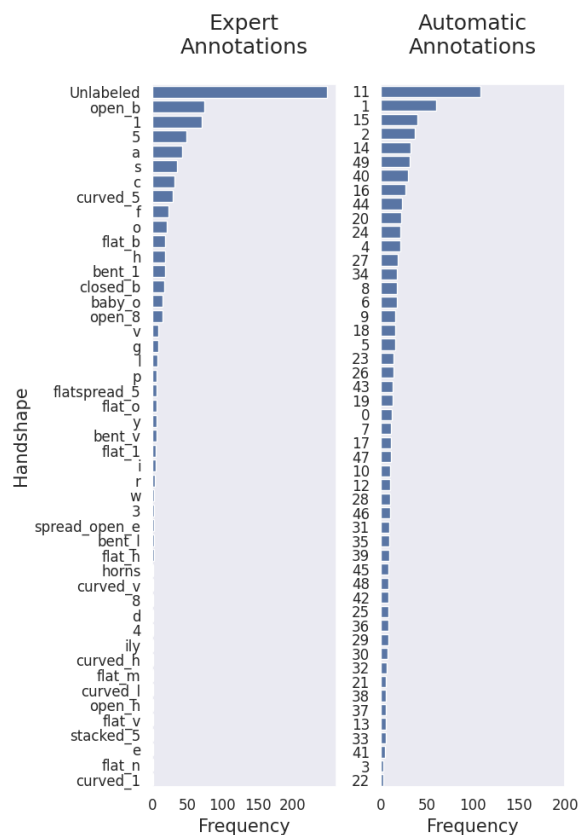


Figure 4: **Distributions of handshapes in the 3D-LEX vocabulary:** the distribution of handshapes as identified by (a) human experts and (b) the automated annotation process described in Section 4.1. The automatic annotations assign arbitrary cluster IDs to different groups of handshapes determined through a K-means clustering method. It’s important to note that these handshape cluster IDs may not directly correspond to the linguistic labels used by human experts in Subfigure 4.a.

Signer ID	01	02	03	04	05
<b>Native Language</b>	NGT	NGT	NGT	NGT	ASL
<b>NGT Signs</b>	10	400	590	0	0
<b>ASL Signs</b>	155	12	0	644	189
<b>Pedal Control</b>	YES	YES	YES	NO	NO

Table 1: **Signer Characteristics:** Native background of each signer and preference for operating (YES) or delegating (NO) the control of the pedal.

**Alignment with existing SL resources** The vocabularies of 3D-LEX have been aligned with existing SL resources to promote research integrating 3D data with datasets comprised of video data and linguistic databases. Table 2 lists the number of glosses in 3D-LEX overlapping with the vocabularies of the aligned resources, the number of sign pose estimations from example videos available for the glosses in the datasets, and the number of glosses that have been provided with expert human annotations for the dominant hand.

The 3D-LEX ASL vocabulary was selected to ensure that a minimum of five reference videos per sign are available in each ASL dataset. Currently, no dataset with multiple reference videos per gloss exists for NGT, but we anticipate that this situation will change in the future. Currently, the SB NGT lexicon (Klomp et al., 2024; Crasborn et al., 2020) provides one reference video for each gloss in the 3D-LEX NGT vocabulary.

	SEMLEX	WLASL	SB NGT
	ASL	ASL	NGT
Vocabulary	1,000	1,000	1,000
Reference Videos	49,274	12,051	1,000
Expert HS	921	695	888

Table 2: **Alignment with other datasets:** The vocabulary overlap, the number of available reference videos, and the number of available expert handshape annotations for the 3D-LEX vocabulary in the SEMLEX, WLASL, and SB NGT datasets.

## 4. Evaluation

To demonstrate the utility of 3D sign data we turn to one of the envisioned benefits mentioned initially: the facilitation of automatic phonetic labeling. In particular, we present a baseline method for semi-automatic handshape annotation. The efficacy of the annotations is evaluated in an ISR task, through comparison with labels provided by linguists and against scenarios devoid of any labels.

While we expect that the data can be used to label other phonetic properties (*e.g.* hand location, movement, orientations, eyebrow position) we here zoom in on the handshapes. This is an intentional choice, as we consider the use of StretchSense gloves to be the most experimental data acquisition technique for sign language capture. The development of semi-automatic annotation methods benefits both linguistic research and various SLP tasks, including recognition and production.

### 4.1. Semi-Automatic Handshape Annotations

In this section, we demonstrate one simple approach for generating phonetic annotations derived from the 3D-LEX handshape data. Due to the absence of an NGT benchmark for isolated sign recognition, we only generate and assess labels derived from the 3D-LEX ASL vocabulary.

Our approach is designed to produce labels that resemble the handshape annotations typically found in ISR benchmarks, facilitating a meaningful comparison between automated and expert annotations. The glosses in ISR benchmarks are commonly assigned a single handshape label, based on the dominant handshape observed in a single reference video. We ensure that the number of possible label classes in our estimations corresponds approximately to the set of classes identified in the video-data benchmark WLASL. For the implementation and instructions on how to replicate our findings, please refer to the GitHub repository.

**Temporal segmentation** To differentiate characteristic handshape signals from any resting or transitional poses, we construct a temporal segmentation method by calculating the Euclidean distance to each frame relative to the calibration poses. This method enables us to perform a first-order discrimination of signals within a recording.

We estimate and segment the poses of both hands to take into consideration that the signer may not strictly enforce the use of their dominant hand. Subsequently, we calculate the frequency of each observed handshape and select the handshape with the highest frame count. As the typically most frequent signal is the resting pose '5', we only select the '5' handshape if it is detected in more than 90% of the frames, otherwise, we select the second most frequently occurring class. The frames where the dominant handshape was detected are then selected as candidate frames for downstream analysis. Figure 5 showcases the output of a Euclidean distance handshape classification approach on frames from the captured sign 'zero'. Here, the handshape 'o' was identified as the characteristic handshape of the sign.

**Semi-automatic labeling** The Euclidean distance labeling technique limits the identification of handshapes to those poses used during the glove calibration phase. This is suboptimal, as the calibration methodology of stretch sensors for capturing sign language is still in a nascent stage. Specifically, the calibration poses may not cover the full range of handshapes present in the lexicons.

To enable a more flexible identification of handshapes, we applied k-means clustering on the average poses of the frames selected during the temporal segmentation. We selected  $k=50$ , which is



### Temporal Segmentation of EDs Poses

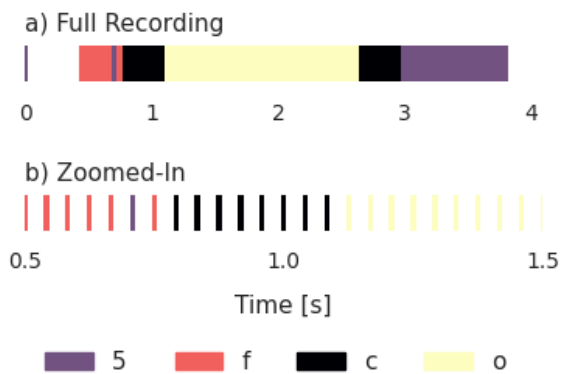


Figure 5: **Time-series visualization of handshape classification:** Classification of the ASL sign 'zero', labeled by experts with the handshape 'o'. Frames are captured and displayed as bars, and each bar's color indicates the handshape, determined by applying the Euclidean distance method frame-by-frame. White space indicates that no data was recorded at that time. The timeline, marked on the x-axis, spans four seconds for this sign. A detailed view at the 1-second mark is provided in the lower row for closer inspection. Our segmentation pipeline identifies the handshapes '5', 'f', 'c', and 'o', selecting frames corresponding to 'o' as the characteristic signal of 'zero'.

approximately the number of handshapes identified in ASL-LEX for the 3D-LEX vocabulary. We assign a new handshape label to each sign in 3D-LEX ASL, corresponding to the arbitrary cluster IDs assigned while clustering the high-dimensional features.

Figure 6 presents a t-SNE projection into two dimensions of the average hand poses, demonstrating that the high-dimensional features cluster. This implies that the signals from the gloves carry sufficient information to distinguish between different handshapes in sign language, revealing distinct characteristics for clusters of signs.

**Evaluation of annotations** To evaluate the efficacy of our annotations, we employed the OpenHands framework (Selvaraj et al., 2022). More precisely, we adopted the framework's adaptation as implemented by Kezar et al. (Kezar et al., 2023), which facilitates gloss recognition supported by phonetic properties. Their foundational work demonstrated that training with phonetic labels enhances gloss recognition accuracy, by merging the WLASL benchmark with the expert linguistic descriptions provided by the ASL-LEX dataset.

In our evaluation process, we trained an SL-GCN (Jiang et al., 2021) architecture to predict glosses within the WLASL dataset, where we use the subset

### t-SNE Projection of Handshapes

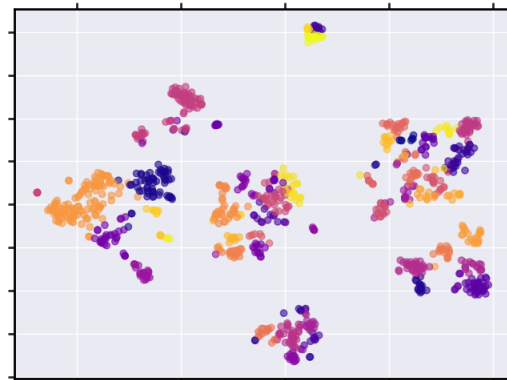


Figure 6: **t-SNE projection:** A t-SNE projection of average hand poses into two dimensions, where the poses were averaged across temporal segments of each sign determined by the Euclidean segmentation method. The projection space lacks units and aims solely to illustrate how high-dimensional 3D-LEX handshape features cluster, highlighting distinguishable signals. Each color represents one of 50 k-means cluster IDs, serving merely to aid visual differentiation of the clusters.

of the WLASL data which overlaps with the 3D-LEX vocabulary. Training persists until validation accuracy ceases to improve for 30 consecutive epochs. The distribution of files across the training, validation, and test splits utilized in our experiments is detailed in Table 3.

To provide a baseline for comparison, we trained the SL-GCN to predict glosses both with and without leveraging handshape labels from the ASL-LEX. Subsequently, we substituted the ASL-LEX handshape labels with our semi-automatic annotations and retrained the models to undertake gloss recognition supported by our annotations. This approach facilitates a comparison of our semi-automatic annotation method against human expert annotations, in terms of their ability to support learning in a downstream ISR task.

Train	Val	Test
8209	2174	1668

Table 3: **Train-Val-Test splits:** Number of examples in the Train-Val-Test splits for the WLASL benchmark experiments.

**Results** The outcome of our isolated sign recognition experiment using semi-automatic handshape labels is presented in Table 4. We provide the top-1 recognition accuracy on the test set, meaning the ratio of how often the model predicted the correct

gloss as the most likely label for a video amongst 1,000 classes. As can be observed, the automatic annotations perform on par with annotations provided by linguistic experts. This is an indication that high-resolution 3D data can offer to reduce the costs associated with linguistic annotation of signs in video datasets and that StretchSense signals are adequate to capture essential handshape features in signs.

$a_1^N$	$a_1^E$	$a_1^A$
0.44 $\pm$ 0.01	0.48 $\pm$ 0.01	<b>0.49<math>\pm</math>0.01</b>

Table 4: **Top-1 recognition accuracy:** Accuracy using no (N) handshape labels, expert (E) labels and automatic (A) labels. The accuracies are averaged across 8 runs, and the standard deviation across measurements is provided in the subscripts.

## 5. Limitations and Prospects

In the process of capturing our data, we have observed many potential areas for improvement. In this section, we highlight some of the current limitations in our methodology, and our intent for addressing them in future work.

Like numerous datasets in sign language research, a significant limitation of 3D-LEX is signer diversity. A dataset comprising a single example for each sign, and which contains only five signers, is insufficient for representing the diversity and rich prosody inherent to sign languages. It is as such not possible to use 3D-LEX in isolation to learn representations useful in sign applications. Consequently, 3D-LEX can primarily serve for limited feature studies or to support video datasets by either providing a 3D ground truth or synthesizing multi-view 2D data from one signer. Future work should consider exploring 3D data which includes both multiple examples per signer and multiple signers per gloss.

While all participants were native signers, it is critical to highlight that only one had ASL as their primary language. As a result, a significant segment of the 3D-LEX ASL dataset was produced by signers whose primary language is NGT but who were proficient in ASL. The impact of employing signers whose primary sign language differs from the captured target language, on the quality and authenticity of lexical sign data remains an area for future research. This concern is recognized as a limitation in v1.0 of 3D-LEX.

The dataset has a limited scope, which comprises a non-exhaustive set of phonological features and vocabularies from the complete languages. However, our method facilitates the pro-

duction of larger vocabularies and data for additional sign languages.

We observed several limitations in our current pipeline. While experimenting with the data acquisition control we noticed varying preferences among signers for operating the pedal. The choice of operator resulted in the emergence of several distinct patterns within the data. When signers themselves operate the pedal, it's generally more efficient but introduces a signal from foot movement at the start and end of each sign. Conversely, using an external operator can result in greater variability in the timing of recordings, affecting the consistency of the recorded time window around each sign. Efforts to streamline these production elements are anticipated in future work.

While our system has been designed with a focus on efficiency, we have identified several limitations concerning the hardware. To the best of our knowledge, 3D-LEX is the first publicly available dataset using the StretchSense gloves to conduct statistical analysis on handshapes in sign language. These gloves were initially developed to generate animation data, which typically does not require the same degree of accuracy in capturing detailed, varied and intricate movements of fingers and hands. Therefore, employing these gloves to provide detailed studies of handshapes in sign language represents a novel and experimental approach. Although the gloves have shown promising capabilities, their performance has presented several challenges.

Notably, the precision of the gloves' measurements is closely tied to how well they fit the signers' hands and the length of time they are worn. A snugger fit typically leads to higher accuracy. However, prolonged usage has been observed to decrease accuracy, likely due to the glove's position shifting on the hand, thereby deviating from its calibrated stance. Shifts can occur for example when hands swell from accumulated heat and from natural movements during wear. Larger gloves relative to the hand size are more prone to positional shifts, exacerbating this issue.

The Hand Engine software is prone to overfitting the sensor data to the calibration poses, a tendency that amplifies when training involves an extensive calibration pose set. Currently, the calibration process utilizes either 20 or 25 poses. We observed that such a detailed pose repertoire complicates Hand Engine's ability to accurately replicate more complex poses and distinguish between poses where the shift in stretching values are relatively small. Figure 7 illustrates a series of poses that exhibit substantial differentiation challenges for the gloves under our calibration framework. With the current version of Hand Engine, future research may gain advantages from employing a smaller set of calibration poses. Ideally, these selected hand-

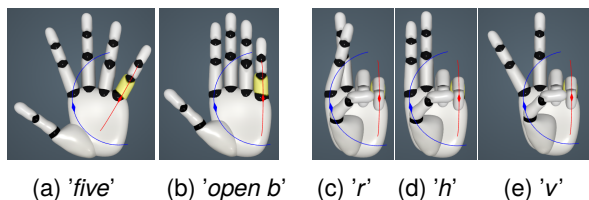


Figure 7: **Failure modes using the StretchSense gloves:** Example handshapes that are challenging to discern for the gloves, conditioned on our calibration scheme. The gloves struggled to differentiate between the handshapes 'five' and 'open b', and between the handshapes 'r', 'h' and 'v'.

shapes should not only be representative of those within the dataset but also exhibit maximum distinction from each other within the calibration set.

An in-depth assessment of calibration methods to address overfitting issues warrants further exploration. This becomes especially critical in capturing continuous signing, where the range of anticipated handshapes is far more variable and unpredictable than in lexical datasets. The 3D-LEX team is actively engaging with StretchSense to enhance glove calibration for sign languages, focusing on better support for continuous signing and capturing a broader spectrum of handshapes. The gloves' ability to accurately represent signing is contingent upon the calibration process, however, as this is a software concern, we expect the conditions for continuous signing to improve in later versions of the Hand Engine software.

Upon assessing the Vicon data, we identified several artifacts occasionally occurring in recordings. For example, we observed random hand orientation flips, which can be attributed to occlusions, where the cameras lost clear line-of-sight to the hands. In such instances the markers may be mistaken for each other, causing the palm to rotate when displayed on an avatar. To mitigate this issue, one can attempt to optimize the positions of the cameras standing on the floor or apply post-processing techniques, such as the filter and gap solver functionalities available in Shogun Post, or by re-labeling the swapped markers.

Moreover, due to limited time, we could not assess the data generated by the LLF application in detail. However, we observed considerable variation in the use of markers like mouthing cues and gaze among participants. In our future research, we aim to delve into these patterns and thoroughly evaluate the quality of the facial feature data.

In our evaluations of 3D-LEX, we presented a basic approach to deriving annotations. However, we emphasize that signs are complex and may contain transitions or oscillate between multiple characteristic handshapes throughout the execution of a sign. While our method approximates the dominant hand-

shape, there are potential benefits in deriving more sophisticated annotation strategies, which consider these transitions and oscillations, and potentially provide multiple phonetic properties for the handshape per sign. However, it is noteworthy that, even in the nascent stages of developing the 3D-LEX production methodology, our automatic annotations yield benefits comparable to those derived from leveraging annotations provided by experts.

## 6. Privacy and Ethical Considerations

The success of machine learning methods has led to large increases in requests for data. While this implies heightened concerns for privacy across computational sciences, it is important to recognize that data collection from minority language communities is at particular risk: Both because a status as deaf classifies as sensitive information, but also because data collection from small populations limits anonymity (Bragg et al., 2020). Additionally, certain sign language datasets that are publicly accessible were compiled without obtaining informed consent from the individuals featured, particularly those datasets that gather information from platforms such as YouTube. All signers contributing to the production of 3D-LEX gave informed consent and received compensation. Moreover, the anonymity of contributors is enhanced compared to typical video datasets, since the motion capture recordings do not visually reveal the signers. To further protect signer anonymity, each participant has been assigned a unique signer ID.

## 7. Conclusion

In this paper, we introduce a new and efficient method for collecting 3D sign language data, resulting in the 3D-LEX dataset, and describe a semi-automatic approach for producing phonetic annotations. The 3D-LEX dataset was produced leveraging three distinct motion capture systems, with two collection techniques to capture manual markers and one technique to capture non-manual markers. Although our approach shows considerable room for improvement, we highlight its potential by automatically generating handshape labels for 1,000 ASL signs. Our initial evaluations of the labels on a downstream ISR task reveal that the semi-automatic annotations offer benefits parallel to those of expert annotations. In conclusion, the 3D-LEX v1.0 demonstrates considerable potential even in its early stages of development. We anticipate that future research using 3D-LEX will investigate synthesizing multi-view data from the 3D ground truths to support tasks such as multi-view SLR, and develop approaches annotating additional phonetic classes.

## 8. Acknowledgments

The team behind 3D-LEX consisted of both Deaf and hearing researchers, whose participation in the project was made possible through financial support from the Platform Digital Infrastructure for the Social Sciences and the Humanities (PDI-SSH) and the Netherlands Organization for Scientific Research (NWO). We extend our gratitude to all external participants who assisted in gathering the data, and to the [Visualisation Lab](#) and [SignLab](#) at the University of Amsterdam for generously providing us with their facilities and equipment. In addition, we wish to thank the company ProCare<sup>9</sup>, who performed the setup of our Vicon rig.

## 9. Bibliographical References

- SK. Ashraf Ali, M. V. D. Prasad, P. Praveen Kumar, and P. V. V. Kishore. 2022. [Deep multi view spatio temporal spectral feature embedding on skeletal sign language videos for recognition](#). *International Journal of Advanced Computer Science and Applications*, 13(4).
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Brafport, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.
- Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2020. [Exploring collection of sign language datasets: Privacy, participation, and model performance](#). In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20, New York, NY, USA. Association for Computing Machinery.
- Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh. 2021. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3413–3423.
- Lee Kezar, Jesse Thomason, and Zed Sehyr. 2023. [Improving sign recognition with phonology](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2732–2737, Dubrovnik, Croatia. Association for Computational Linguistics.
- Oscar Koller. 2020. [Quantitative survey of the state of the art in sign language recognition](#). *ArXiv*, abs/2008.09918.
- Gomèr Otterspeer, Ulrika Klomp, and Floris Roelofsen. 2024. Signcollect - a 'touchless' pipeline for constructing large-scale sign language repositories. In *LREC-COLING 2024 - 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, Turin, Italy.
- Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, and Mohammad Sabokrou. 2021. Sign language production: A review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3451–3461.
- Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. [OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133, Dublin, Ireland. Association for Computational Linguistics.
- William C. Stokoe. [Sign language structure: An outline of the visual communication systems of the american deaf](#). *The Journal of Deaf Studies and Deaf Education*, 10(1):3–37.
- Freya Watkins, Diar Abdikarim, Bodo Winter, and Robin L. Thompson. 2024. [Viewing angle matters in british sign language processing](#). *Scientific Reports*, 14(1).
- Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. [Including signed languages in natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

---

<sup>9</sup>ProCare BV, The Netherlands



## 10. Language Resource References

### Language Resources

- Athitsos, Vassilis and Neidle, Carol and Sclaroff, Stan and Nash, Joan and Stefan, Alexandra and Quan Yuan and Thangali, Ashwin. 2008. *The American Sign Language Lexicon Video Dataset*. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.
- Benchiheub, Mohamed-El-Fatah and Berret, Bastien and Braffort, Annelies. 2016. *Collecting and Analysing a Motion-Capture Corpus of French Sign Language*. Workshop on the Representation and Processing of Sign Languages.
- Helen Cooper and Eng-Jon Ong and Nicolas Pugeault and Richard Bowden. 2012. *Sign Language Recognition using Sub-Units*. Journal of Machine Learning Research.
- Crasborn, Onno and Bank, Richard and Zwitserlood, Inge and van der Kooij, Els and Ormel, Ellen and Ros, Johan and Schüller, Anique and de Meijer, Anne and van Zuilen, Merel and Nauta, Yassine Ellen and van Winsum, Frouke and Vonk, Max. 2020. *NGT dataset in Global Signbank*. Radboud University, Centre for Language Studies, ISLRN 976-021-358-388-6.
- Amanda Cardoso Duarte and Shruti Palaskar and Deepti Ghadiyaram and Kenneth DeHaan and Florian Metze and Jordi Torres and Xavier Giró-i-Nieto. 2020. *How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language*. ISLRN 583-408-694-292-6.
- Gao, Liqing and Zhu, Lei and Xue, Senhua and Wan, Liang and Li, Ping and Feng, Wei. 2023. *Multi-View Fusion for Sign Language Recognition through Knowledge Transfer Learning*. Association for Computing Machinery, VRCAI '22.
- Gibet, Sylvie. 2018. *Building French Sign Language Motion Capture Corpora for Signing Avatars*.
- Heloir, Alexis and Gibet, Sylvie and Multon, Franck and Courty, Nicolas. 2006. *Captured Motion Data Processing for Real Time Synthesis of Sign Language*. Springer Berlin Heidelberg.
- Huang, Jie and Zhou, Wengang and Zhang, Qilin and Li, Houqiang and Li, Weiping. 2018. *Video-based sign language recognition without temporal segmentation*. AAAI Press, AAAI'18/IAAI'18/EAAI'18.
- Jedlička, Pavel and Krňoul, Zdeněk and Kanis, Jakub and Železný, Miloš. 2020. *Sign Language Motion Capture Dataset for Data-driven Synthesis*. European Language Resources Association (ELRA).
- Hamid Reza Vaezi Joze and Oscar Koller. 2019. *MS-ASL: A Large-Scale Data Set and Benchmark for Understanding American Sign Language*. BMVA Press.
- Kezar, Lee and Thomason, Jesse and Caselli, Naomi and Sehyr, Zed and Pontecorvo, Elana. 2023. *The Sem-Lex Benchmark: Modeling ASL Signs and their Phonemes*. ACM, ASSETS '23.
- Klomp, Ulrika and Gierman, Lisa and Nauta, Ellen and Otterspeer, Gomèr and Pelupessay, Ray and Stern, Galya and Wubbolts, Casper and Oomen, Marloes and Roelofsen, Floris. 2024. *An extension of the NGT dataset in Global Signbank*. SignBank.
- Kopf, Maria and Schulder, Marc and Hanke, Thomas. 2022. *The Sign Language Dataset Compendium: Creating an Overview of Digital Linguistic Resources*. European Language Resources Association.
- Li, Dongxu and Rodriguez, Cristian and Yu, Xin and Li, Hongdong. 2020. *Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison*. The IEEE Winter Conference on Applications of Computer Vision.
- Lu, Pengfei and Huenerfauth, Matt. 2010. *Collecting a Motion-Capture Corpus of American Sign Language for Data-Driven Generation Research*. Association for Computational Linguistics.
- Mopidevi, Suneetha and Prasad, M.V.D. and Kishore, P.V.V. 2023. *Multiview meta-metric learning for sign language recognition using triplet loss embeddings*. Pattern Analysis and Applications.
- Mariusz Oszust and Marian Wysocki. 2013. *Polish sign language words recognition with Kinect*. 2013 6th International Conference on Human System Interactions (HSI).
- Razieh Rastgoo and Kourosh Kiani and Sergio Escalera. 2020. *Hand sign language recognition using multi-view hand skeleton*. Expert Syst. Appl.
- Adam C. Schembri and Jordan B Fenlon and Ramas Rentelis and Sally Reynolds and Kearsy Cormier. 2013. *Building the British Sign Language Corpus*. University of Hawaii Press.
- Sehyr, Zed Sevcikova and Caselli, Naomi and Cohen-Goldberg, Ariel M and Emmorey, Karen. 2021. *The ASL-LEX 2.0 Project: A Database*

*of Lexical and Phonological Properties for 2,723 Signs in American Sign Language.* The Journal of Deaf Studies and Deaf Education.

von Agris, Ulrich and Kraiss, Karl-Friedrich. 2010. *SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition.* European Language Resources Association (ELRA).

# Signbank 2.0 of Sign Languages: Easy to Administer, Easy to Use, Easy to Share

Ronice Muller de Quadros<sup>1</sup> , Christian Rathmann<sup>2</sup> , Péter Zalán Romanek<sup>2</sup> ,  
Francisco Fernandes<sup>3</sup> , Sther Condé<sup>3</sup> 

Universidade Federal de Santa Catarina<sup>1</sup>, Humboldt-Universität zu Berlin<sup>2</sup>, Levante Lab<sup>3</sup>  
[ronice.quadros@ufsc.br](mailto:ronice.quadros@ufsc.br), [christian.rathmann@hu-berlin.de](mailto:christian.rathmann@hu-berlin.de), [peterzalan.romanek@gmail.com](mailto:peterzalan.romanek@gmail.com),  
[francisco.fernandes@levantelab.com.br](mailto:francisco.fernandes@levantelab.com.br), [sther.conde@hotmail.com](mailto:sther.conde@hotmail.com)

## Abstract

Signbank 2.0 integrates sign language documentation to identify signs with their specifications in the context of a large sign language corpus. Signbank 2.0 is inspired by Global Signbank, especially with respect to the integration of the general linguistic structure, and by developments from the earlier Libras Sign Identification platform, with search systems organized by sign language parameters. The current proposal presents several advances, especially regarding the administration panel with a simple dashboard. In addition, the current Signbank 2.0 implements [and at least one more instance] more sophisticated search systems from a linguistic and technological point of view. The tools developed include more possibilities for sign searches categorized based on linguistic and visual criteria. Finally, the search system presents the frequency of signs linked to the EAF files, listing the occurrences in the integrated corpus and giving the exact video timing of the sign.

**Keywords:** Signbank, Sign language documentation, Sign language visualization, Visual design

## 1. Introduction

Signbank 2.0 is a database of signs from different sign languages associated with corpora. It is the result of previous sign databases developed with the aim of providing descriptions of each sign as a list of signs extracted from a specific sign language corpus. Johnston (1989) created the first lexical database for sign language. His work aimed to provide a dictionary of Australian Sign Language (Auslan) based on the Auslan corpus. Johnston's work led to the establishment of the Global Signbank (Cassidy et al., 2018; Crasborn et al., 2012, 2018). The Global Signbank was an initiative to create a global database of different sign languages. In parallel, Brazil created the first Sign ID for Brazilian Sign Language (Libras) in 2008 (Quadros, 2016; Quadros et al., 2020). This specific Sign ID had the basic goal of listing the signs associated with the gloss identification words from Portuguese for utility purposes only; the glosses allowed annotators to be consistent in their Libras annotations. Each sign had an associated ID gloss to feed the annotations of the Libras corpus. The Sign ID system also developed a search tool based on sign parameters such as handshapes, and locations. However, this system was not user-friendly. In 2014, the proposal was replaced by the Libras Signbank, inspired by the Global Signbank and used as an open-access system. However, some sign language tools were lost, and the management platform was not accessible to the sign researchers. In 2019, we decided to improve the Signbank with a different system, with new open access software, to include again sign language search tools inspired by the Sign ID system, combined with new developments, subsequently published by Scolari (2022) and Quadros et al. (2022). These new search tools

offer different ways to locate the signs, taking into account general users and users who do not know the glosses that identify the signs. This was done by integrating a sign language-based search tool that starts from the handshapes and includes the hands involved in the sign (one-handed signs or two-handed symmetrical/asymmetrical signs) and the location of the sign (head, torso, limbs).

Moreover, the dashboard has been developed to be accessible to sign language researchers. It is designed for sign language communities, especially deaf communities. The main approach is to decentralize the management of the system, giving the users the right to manage it. The basic idea is “*they can do it themselves*”. This dashboard contains tools that are sophisticated but easy to use and accessible to every member with different roles in the system. The roles created include (i) ‘Administrator’, (ii) ‘Data Publisher’, and (iii) ‘Data Publication Approver’. The administrators can manage the organization, the data and the categories integrated in the Signbank 2.0. This design was done by the developers, who reviewed the users' workflow and fed it into the creation of the solutions.

Signbank 2.0 is currently being tested for Brazilian Sign Language (Libras) (<https://signbank.libras.ufsc.br/en>) and will soon be available for other sign languages (International Sign Lg., IntSL), German Sign Language, DGS), Hungarian Sign Language, Austrian SL, and Estonian SL, with the possibility to be applied to other sign languages over the world for parallel analysis through a next step development that will possibility the network among all signbanks 2.0. As an example, the Libras Signbank contains 3,067 signs with image,

video, and phonological descriptions that allow searching by handshape, location, and handedness (one-handed signs, symmetrical/asymmetrical two-handed signs). The information about phonological features is inserted with codes and handshape images. The option for handshape images is preferred by users because it is known to them and easily identifiable. Both administrator and general users have access to the handshape images for both hands. The codes associated with the handshapes follow the Global Signbank with some adjustments. The handshape search tool accessed by general users can be associated with either HamNoSys or SignWriting in the next stages of the database. Currently, the search tools are based on linguistic descriptions selected from lists using the written form and images of the handshapes and icons/symbols. These choices are related to previous experience with older versions of the Signbank where we used SignWriting. Users, including deaf users, did not use it as a reference for their search. In fact, they used various guesses of possible written words to try to find the sign, or they signed in specific groups using social media tools to find out what the gloss was for the particular sign they needed to annotate. This user experience/feedback led to the development of the handshape slider by the design student at the Universidade Federal de Santa Catarina. This slider was incorporated in Signbank 2.0 (Scolari, 2022; Scolari et al. 2022).

Signbank 2.0 is a technology-mediated collaborative environment that meets Davidson's (2008) definition of a generation of tools called Humanities 2.0, in which participation is based on different sets of theoretical assumptions that decenter knowledge and authority. The foundation of the Signbank's current structure is based on a community of practice that benefits from technologies to amplify the networks of relationships, making learning and social construction of knowledge possible through creative techniques and the use of tools (Wenger et al., 2002, see also Quadros et al., 2022 for Libras).

The technical architecture of the Signbank consists of a systematic and structured approach to designing and defining the structure, components, and interactions of a complex system. The requirements for the development were meticulously carried out through a series of immersion phases derived from the participatory design methodology (Camargo & Fazani, 2014). This collaborative approach involved stakeholders and end users actively involved in the development process, ensuring that their perspectives and needs were thoroughly considered. Through meetings, interviews, and iterative feedback loops, we gained valuable insights that shaped the project's direction, resulting in a user-centered and highly effective

solution that precisely meets the expectations and requirements of its intended users. This project, focused on the coexistence of sign language and deaf communities, has led to the development of Signbank 2.0.

Understanding the needs of the users, the niche and the public is the first step in this process. This was done by interviewing stakeholders, conducting scenario and user research, and defining and systematizing common platform requirements. The next step was to analyze and synthesize the results and draw some conclusions. This was done by drawing conclusions and synthesizing the research, developing personas and a User Journey Strategy as a procedural strategy for Thinking Design. The next step was to create, prototype, and test. With well-defined strategies in place, the path was clear to create all the necessary pieces to execute the project. Our focus was on designing the ideal solution, using collaborative creation (co-creation) and evaluation tools to help with this process. Then we have the style guide, site map, prototyping (low, medium and high fidelity) and usability tests. Once the tests had been completed, the development of the scripts began. The prototype was mature enough to be implemented, allowing programmers to code and give materiality to the project.

User feedback was collected in a system designed for interaction between users and developers along the process. It was designed as a collaborative form where users review each step of development and add suggestions when needed. The basic idea was to make the communication between users and developers very efficient, because in previous experiences with the development of previous versions of the signature bank, we learned that this is a key step in the process.

The evolution of Signbank 2.0 allows users to have autonomy to manage the system. It removes the barriers imposed by the limitation of language specificity and allows the modification of sign-related features. Thus, Signbank 2.0 has a structure that can be replicated by different institutions and adapted to different sign languages and countries. Our goal was to provide a sign language documentation tool that could be used by sign language communities and research communities, creating opportunities to create a Signbank in their own countries, especially those with limited financial resources.

Considering the target audience of the Signbank and the needs of sign language communities and research communities (including deaf and non-deaf researchers), Signbank 2.0 was designed to include aspects related to web accessibility, usability, and visual organization. The main goal was to have a platform that was friendly to signers



(i.e., not only to computer technicians) and easy to manage, use, and share. A complete set of signed videos explaining each page was created. The administrators can edit these videos of the pages at any time. The organization uses videos available in sign language corpora associated with EAF files from ELAN Eudico Annotator (Crasborn et al., 2012). These sign language corpora feed the Signbank, which complement the signs with specific linguistic information. Another important aspect is that the current Signbank 2.0 is designed to be sustainable considering its technological lifetime and version developments. The main sustainable goal is that the community of users at universities and research institutes worldwide will continue to improve it technically by implementing a multicenter Signbank 2.0. network.

The development of this research and the resources for accessibility are described in this article.

The architectural basis of Signbank 2.0 allows its application to sign languages in other countries. As a result, documentation is available for Brazilian Sign Language (Libras), German Sign Language (DGS), International Sign Language (IntSL), Hungarian Sign Language (MJNY) and Austrian Sign Language (ÖGS). It is open-source software with the goal of making it sustainable through network platforms to be implemented in the next steps connecting all the signbanks of the sign languages that have implemented it.

Signbank 2.0 is a linguistic corpus-based tool, not a bilingual dictionary. The motivation for the first versions of signbanks around the world was related to the need to have standard glosses to identify signs, so we refer to the glosses as ID-gloss or ID-sign. However, considering the development of sign language corpora all over the world, the signbank started to include corpus-related information that identifies each sign based on linguistic information (such as phonological, morphological, syntactic and semantic, and more recently iconicity), expanding the original concept. Signbank 2.0 contains all this technical information and possible translations for each sign. The possible translations also serve sign language annotation purposes, as annotation can include the translation of sign production into another language.

## 2. Resources to Signbank 2.0

Signbank 2.0. has two basic interfaces available to its users: (a) the portal and (b) the dashboard. The portal is available to all users who want to access the database for various purposes, e.g. to find a specific sign, to view the occurrences of the sign in the available corpus, to identify the glosses associated with a sign, and to research signs for various linguistic and translation purposes. The

dashboard is intended for users with specific roles in Signbank 2.0 (administrators, publishers, and approvers of specific changes). This development gives more control to the end users, as it was built to give them autonomy, independent of the developers.

The tools developed for this new version of Signbank allow the management of resources, including an accessible structure based on sign languages. The background idea is to have a simple but robust platform that can accommodate all the requirements of the Signbank. This follows Rosenfeld, Morville, and Arango's (2015) proposal for building platforms based on the organization of tools that prioritize a layout based on clarity with an architecture using a distribution of information with little depth. That is, only a few clicks are required to access any content of the Signbank 2.0.

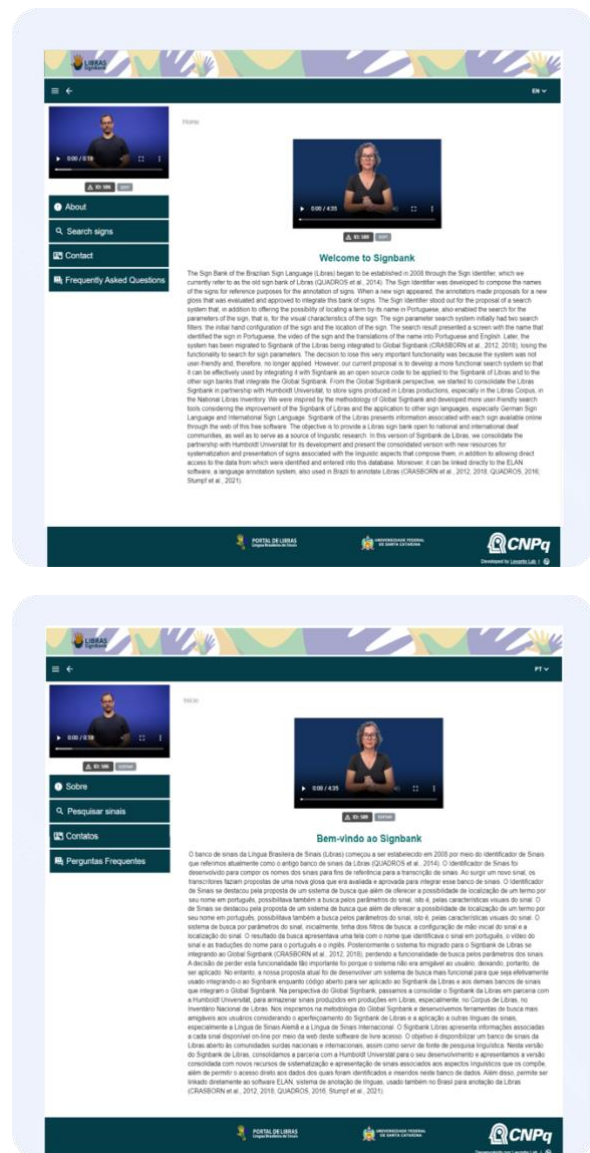


Figure 1: Libras Signbank Landing Page in Portuguese and English.

Source: <https://signbank.libras.ufsc.br/pt/about>

## 2.1 Signbank 2.0 Portal<sup>1</sup>

The Signbank 2.0 Portal contains the following resources:

- A landing page which includes general information about the system and the general layout of the sign bank.
- Search tools with features including handshapes, locations, words, linguistic information and visual network
- Frequency of signs in the sign language corpus
- A list of sign occurrences in the current corpus
- Language contact with sign language(s) and/or written systems
- Frequently Asked Questions (FAQ)
- Terms of Use
- Privacy Policy

The Signbank 2.0 Portal has the information available in the sign language of each country. It can also be accessed in the written language of the respective country and/or in English. The portal layout includes the menu to access the general information on the 'About' page, search tools, contact information and FAQ. Each menu item is accompanied by a signed explanation (see figure 1).

The FAQ, the Terms of Service and the Privacy Policy can be edited in the dashboard as often as necessary, according to legal requirements of the respective country.

The search tools are an innovative part of Signbank 2.0. Despite their complexity, they are designed to be intuitive and comprehensible and to be used in different ways. Figure 2 shows the options to search for signs:

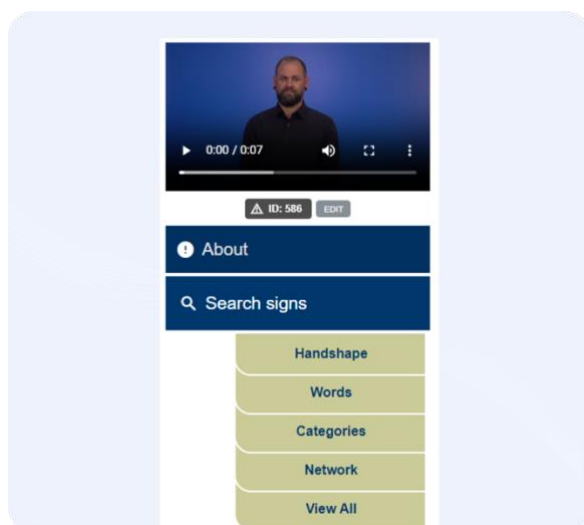


Figure 2: Signbank search tools menu

The sign search by handshape is the result of research by Scolari et al. (2022). This is a new design which is considered a novel solution to the problem identified in the Sign ID search system. The order of the handshapes is organized based on similarity organization. This search tool allows the users to scroll easily through all the handshapes listed in each sign language, as illustrated in Figure 3, using a scroll bar.

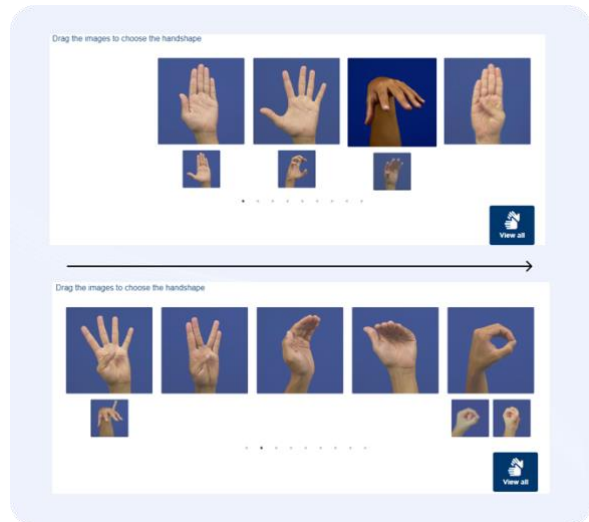


Figure 3: Visualization of the handshape scroller

A major improvement over the previous systems, which was/were not user friendly, is that the search tool in Signbank 2.0. allows users to scroll quickly through all the options on the same page. The previous system grouped handshapes and changed pages for each group. User would become lost among all the options, and it was complicated to reload the pages to find the option to select. Signbank 2.0 has all the handshapes on the same line, so users can scroll forward and back easily to find the exact option that fits the sign they are searching for.

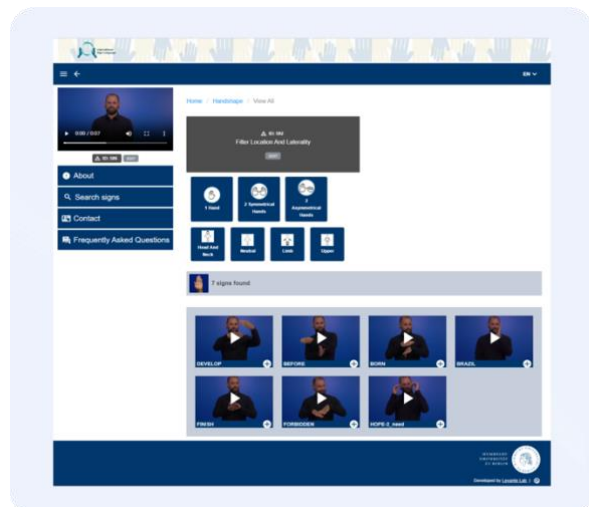


Figure 4: Additional Filters for Searching Signs

<sup>1</sup> Signbank 2.0 functionalities are listed in the appendix.

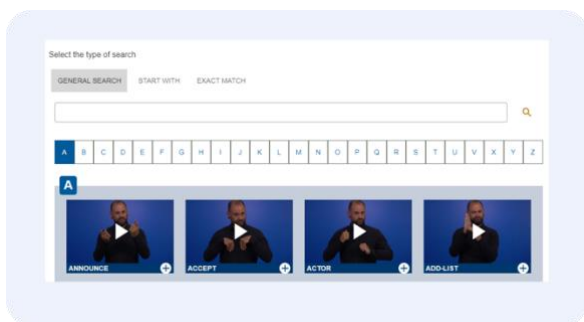


Figure 5: Word-based search for signs

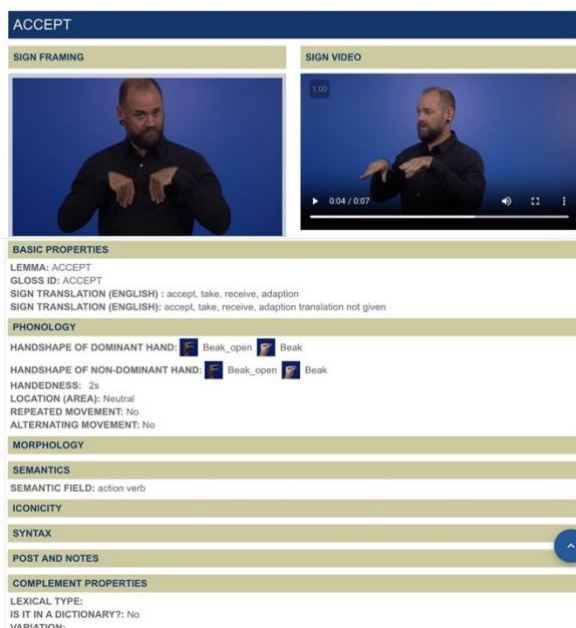


Figure 6: Results of the exact sign with a list of properties if available

In addition, the less common variants of the handshape are located directly below the main handshape, so that all handshapes can be displayed in an easily recognizable visual size. Usability tests were conducted with users of the Libras Signbank. The results indicate that users take advantage of finding the signs using this new search tool. The search tool has additional filters that include the number of hands and their arrangement involved in the sign and the location where the sign is typically produced. Figure 4 illustrates these additional filters.

These filters include options to restrict the search by the number of hands involved in the sign (one-handed signs vs. two-handed symmetrical signs and/or two-handed asymmetrical signs). The features of location in the search tool are a) around the head, b) neutral space, c) limbs and d) upper torso. Another search tool is word-based, as shown in Figure 5.

In this case, it is possible to use an initial letter that the gloss ID starts with. Alternatively, it is possible to search by choosing from the options: general search, start with an exact word. These options

are designed to serve different purposes. Annotators usually do not know the gloss ID when they are looking for it to follow the standard annotation of a particular sign. Thus, they may have clues about possible words and use a general search to get a list of all the tags that use a possible word, and then look at the tag listed. Sometimes they remember the first letters of the word and choose the second option. If we know the exact gloss ID for a sign, we may want to search for it directly to get the list of occurrences for research purposes. In this case, it is possible to look at each occurrence directly in the corpus with the full list of places where it appears (see Figures 6 and 7).

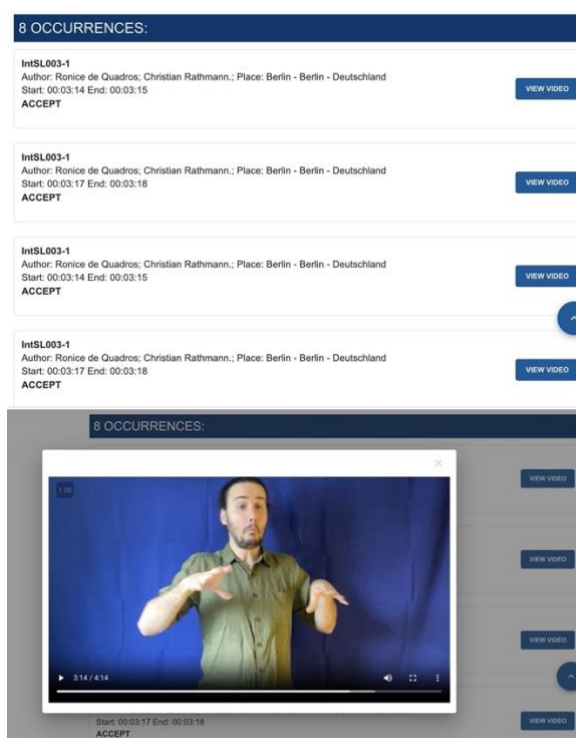


Figure 7: List of occurrences of the sign ACCEPT in the current corpus & visualization of one of the available occurrences

Each occurrence can be accessed directly at the exact time in the video where it occurs. This is made possible by the EAF files associated with the videos in the corpus. In the case of the International Sign Language (IntSL) Signbank shown as an example in these figures, the EAF files are annotated in English, which is the only language available to date. However, for national Signbanks, such as for Libras, for DGS, for MUNY, for ÖGS, there are two ways to search for signs: by gloss ID in the national language, such as Portuguese, German or Hungarian in these respective Signbanks, or by English gloss ID. For these two options, we have EAF files annotated in both written languages.



The other search tool is based on linguistic categories: phonology (dominant hand, weak hand, location, movement, orientation, relationship between manual articulators), morphosyntax (word classes), semantics (semantic fields), and complementary properties (variation). These categories can be listed to show all the information about the sign, as shown in Figure 8. The user can choose what to compare between signs in this search option. It is also possible to download the search result in an Excel file.

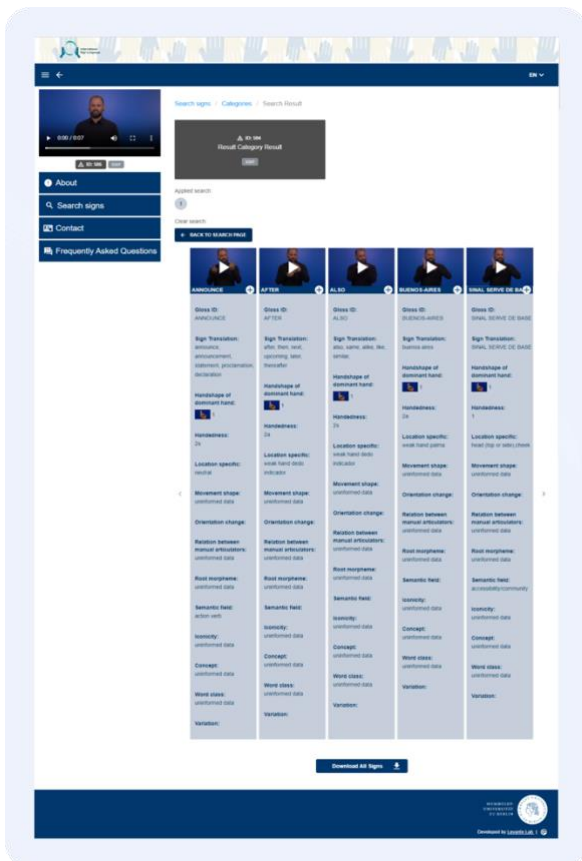


Figure 8: Sign with linguistic information filtered by category of 'dominant handshape'

The network search in turn generates results that include all signs or select linguistic categories in the form of a word cloud. The result of all signs available in Signbank 2.0 shows the signs with more occurrences in larger letters than those with fewer occurrences, as illustrated in Figure 9 for signs in the International Sign Language Signbank.

The network search generates results that show all signs or selected linguistic categories. The result of all signs available in Signbank 2.0 shows the signs with more occurrences larger than those with fewer occurrences, as illustrated in Figure 9 for signs in the IntSL Signbank.

In Figure 9, the signs such as ACQUIRE, BUT-2, and ALREADY show a high frequency of occurrences in the IntSL corpus, which is

associated with the IntSL Signbank. On the other hand, the signs with smaller word sizes placed in the network visualization are the ones with lower frequency of occurrences. For example, CURRICULUM has 2 occurrences. ACQUIRE has 140 occurrences and BUT-2 has 116 occurrences in the IntSL corpus. A slightly larger word, such as BOOK, has 16 occurrences, and the other word even larger than BOOK, such as CLEAR, has 42 occurrences in this corpus.



Figure 9: Network search results in the IntSL Signbank

The last option is to display all the signs. This is useful because annotators sometimes want to look at the whole set of signs. It was noticed in the Sign ID system that deaf annotators used to ask for administrative access in order to access all signs in the Sign ID. Based on this experience, this option has been added to the Signbank 2.0, as shown in Figure 10.

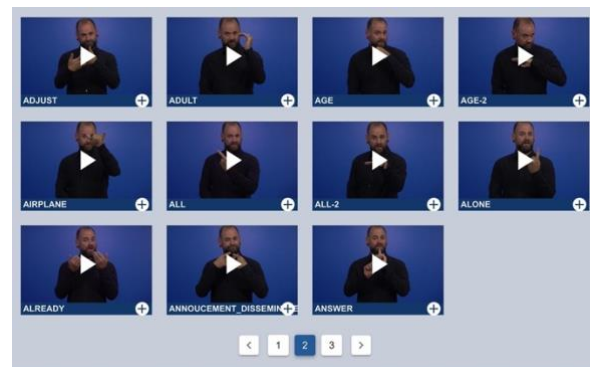


Figure 10: Paged display of available signs with the scroll to see the next signs



Figure 11: Options for searching signs



Users can also search by the Roman alphabet system or by handshapes on the general sign search page (see Figure 11).

Additionally, the portal contains the section of FAQs about Signbank 2.0 in sign languages and written languages. There is also a contact section where users can send direct messages to the system administrators, either in sign language or in written language.

Overall, the creation of Signbank 2.0 is the result of research in the field of design for the development of a visual identity project that values visuality, visual sign language(s) and the forms of visual orientation of sign language users. In addition to adopting the guidelines of web accessibility, it follows the recommendations of studies which analyzed the use of web environments by deaf people, (Flor, 2016 and Fajardo, Parra, and Cañas, 2010), see also the design of the Libras portal in Quadros et al. 2022). The basis of these recommendations always considers the use of visual sign language(s) and contextualized visual resources. These designs privilege the use of familiar and iconic images inspired by specific sign languages to facilitate the understanding of sign language users. The interface has been produced from a deaf perspective, relying on deaf sign language users and sign language researchers throughout its development. Signbank 2.0 takes into account these requirements and includes navigation tools with visual and sign orientation. It is relevant to address that these visual tools are among the top results for accessibility and friendly database use by general users. The possible addition of notation systems, such as HamNoSys and SignWriting, would be for more technical users, for translation purposes, and for the inclusion of avatars in the system, which we are leading for future developments.

## 2.2 Signbank 2.0 Dashboard

The Signbank 2.0 Dashboard is designed to empower the administrative users who manage this portal. It is designed for users to adapt and customize the information needed in each research institution, according to the respective sign language, visual identity of the platform, about, terms of use and privacy requirements of each country. The administrators of the research institution can manage all this specific information in their Signbank of the respective sign language. They can also customize the specific information about the sign language, such as the sign language categories, including handshapes. The basic idea was to have the ability to feed Signbank 2.0 at any time and make adjustments as needed, without developer involvement. The proposal was to create a dashboard in a simple way for the managers who are allowed to make changes in this portal. This required the definition of “persons”, which includes manager roles with

different tasks. It is also necessary to use the sign language of each country as one of the main languages of the portal to provide all the dashboard information. This makes the Signbank accessible for sign language users.

The Signbank 2.0 Dashboard includes all the settings that can be managed by the users, as shown in Figure 12.

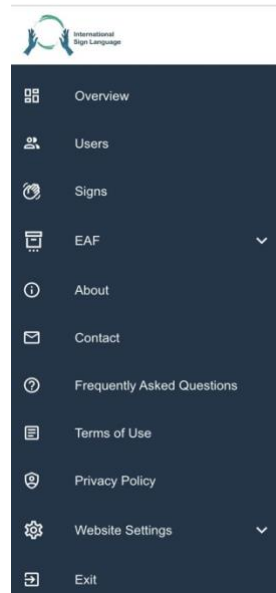


Figure 12: Signbank 2.0 Dashboard general menu

From here, it is possible to manage the whole of Signbank 2.0. People with different roles can make changes or updates in this dashboard. The dashboard can have users with different permissions. They can be administrators, data publishers and data approvers. Administrators can enable or disable approvers or publishers. Approvers can also publish signs, in addition to approving what the publisher has uploaded and filling out any sign included in Signbank 2.0.

In the Sign menu, the administrator can edit the linguistic specifications and there is a list of sign items. Also, publishers can download a new sign and add its specific information, and approvers can approve the signs published by the publisher. Figure 13 shows the view of this area:

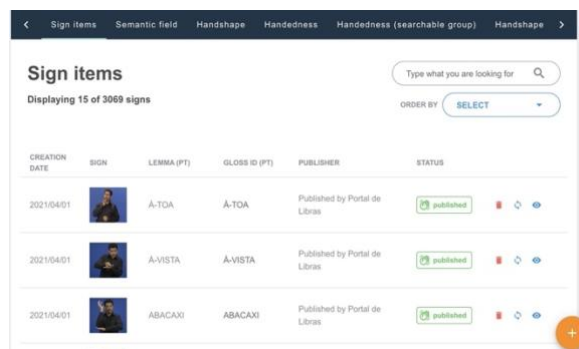


Figure 13: Sign items and linguistic specifications

The top row contains the linguistic specifications that the administrator can edit. The 'Sign Items' list contains all the published signs and their status: they can be approved or pending. Approvers need to check them to approve or to label as 'pending'. To publish a new sign, users click on the + sign at the bottom right of the page. This will open the form to be filled in. This is where the publisher or user with a higher level of access adds all the required and available information, as well as additional information, adding the video signing the sign and the cover with the frame that can better identify the sign. The first page contains the required information that the search tools will use. The following pages contain additional information that is optional and may or may not be used for search purposes, depending on its availability.

The next item on the menu is 'EAF', where the publisher adds EAF files and corresponding movies from the available corpus. When the EAF files and the videos have been included, it shows the list of published EAF files and the list of occurrences of each sign. Figure 15 shows the latter list. Dashboard managers can then visualize all published materials.

CREATION DATE	VIDEO	GLOSS ID	TIME	LANGUAGE
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	GOOD	00:00:04 - 00:00:04	en
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	EI(POSITIVE)	00:00:04 - 00:00:05	en
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	LIKE	00:00:05 - 00:00:06	en
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	POSS(my)	00:00:07 - 00:00:07	en
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	NAME	00:00:07 - 00:00:08	en
2023/07/04	FLN_GR_F06_entrevista_cam03 translation not given (Florianoópolis - Santa Catarina) Ronice Müller de Quadros	F3(sedna)	00:00:08 - 00:00:11	en

Figure 15: List of the occurrences of each sign item available in the Signbank 2.0

The sections 'About' menu, 'Contact', 'Frequently Asked Questions', 'Terms of Use', and 'Privacy Policy' can be updated whenever necessary. For each of them, there is a list of previously published versions, the current one, and the possibility of adding new versions. After the version of the 'Privacy Policy' or 'Terms of Use' has been revised, users will be able to read and accept the updated version.

The last menu entry is that of the site settings. This includes a submenu for languages, institutions, professions, manuals, instructional videos, and platform identity. The manuals include updated versions of the platform manual and the annotation manual. These manuals can be updated to the latest versions, but previous versions are available for reference. Users in the platform access these latest versions of the manuals. The visual identity of the platform can also be modified as needed in the 'Platform Identity' submenu.

The publisher in the portal can visualize all changes to the dashboard. The edit history is listed and can be located using search tools within the dashboard. The whole system is designed for easy visualization and editing. The administrator can manage the roles of managers and the whole dashboard. It is important to note that developers work together with users, discussing and testing all implementation steps. We started with several meetings to understand all the requirements of Signbank 2.0, then designers prototyped the whole system for users to evaluate before developers started to produce the platform. The whole process is planned in a participative construction with all the actors: computer science engineers, designers, manager users and end users.

### 2.3 User evaluation

For the user evaluation, a workshop was organized with a small number of future users to evaluate the interface of Signbank 2.0 and its usability. The feedback of the users is overall positive, and they addressed a few topics.

Firstly, the Signbank 2.0 is also user-friendly for linguists and non-linguists. Persons who are not linguists can use it easily yet can access complex information about the existing lexical items. Annotators with basic linguistic knowledge can upload annotation files and videos and fill in the lexical information in a few steps. Explanatory videos in sign language guide the users as part of the user manual. Secondly, the users appreciated that Signbank 2.0 can read different annotation templates from other sign language corpora with modifications on files. Signbank 2.0 needs only an ID gloss tier to read the tokenized signs; thus, video-recorded materials with annotated files from different everyday language settings can be uploaded into Signbank 2.0. It allows us to expand the set of growing natural data that will be read by Signbank 2.0. Furthermore, the users will get contextual linguistic information, too, because Signbank 2.0 shows the appearance of certain lexical items by displaying the uploaded videos within the range of sign appearance. It is an advantage for different users like linguists, educators, trainers, students, and learners to see the sign in their natural contexts.

Thirdly, the users found the Signbank 2.0 interface is clear, yet the search engine is slightly complex. Persons without linguistic backgrounds may use the search engine with difficulties. However, it needs only three or four clicks to find any signs with the search parameters. They thought the sign frame (picture) was too small to present the salient form of signs. Fourth point: the users considered Signbank 2.0 a good toolkit for the verification process of the registered lexical items based on existing natural data because it is data-driven. The lexical items of Signbank 2.0 will

emerge from the context of natural linguistic behavior embedded in the uploaded videos via the glossing/lemmatization process.

The main problems identified were related to the search tools and the frequency of signs. The results were not correct because the system did not search in the appropriate level reference of the corpus.

### 3. Final Considerations

The development of Signbank 2.0 is the result of the experience with the Sign ID glosses and previous versions of the Signbank, starting with the technology available at the time of implementation in 2008, and the experiences of developers, administrators and end users. The identified problems with the design of the previous platform, with the search tools, and with the management of the changes required allowed us to design and build this robust platform with the portal and the dashboard.

The Signbank 2.0. is a sign language database that is mainly accessed by sign language professionals or students conducting studies with sign languages. It is also used by translators and interpreters to check the translations that a sign may have. However, the most common use is for technical reasons, when annotators are working with annotations and need to know the standard gloss associated with a particular sign, or when researchers are analyzing signs in different contexts of sign production. Interestingly, other signers also access the signbank to review signs for learning purposes. This includes both deaf and hearing people. In general, deaf signers appreciate the search options, as they include different visual representations of the results (word clouds, lists of signs side by side depending on linguistic features, and all signs based on specific selections). Non-signers can also use the database because there are options based on searching by letters or words. However, we have seen that deaf and hearing signers are the main users.

The user with the administrator role can manage the system tools. For example, they can add new handshapes to the list of available handshapes; they can add linguistic information to be filled when a sign is added to the database; they can change tutorials, they can update condition terms; they can change the logo, color patterns, fonts, instruction videos, menu videos, tutorial videos, web texts.

Signbank 2.0 is being developed to be applied to multiple sign languages in parallel, possibly building a sustainable network between the different sign languages. It is important to clarify that the Global SignBank concept has been adopted to develop SignBank 2.0. The move to a new version of this system is related to using new

systems available considering open access tools incorporated into the Signbank 2.0. The architecture of the applications that make up the platform uses the PHP language for the backend application (from the LARAVEL framework) and JavaScript for the frontend application (from the VUEjs framework). Communication between applications will be structured using the REST standard. Database default is structured with MYSQL. The evaluation process of the Signbank is happening along the development process through a collaborative design with deaf users and hearing signers related to sign language studies. The goal was to make available search tools and sign language data in different ways for different purposes, such as finding a written standard identity, visualizing the signs of specific linguistic categories, visualizing the frequency of the signs in the corpus available, searching signs by handshape, hands used in the sign, and location, visualizing the clouds of signs in the system with the possibility to restrict the linguistic category. The design was developed with visual design in mind for deaf people. The prominent target people are deaf and hearing signers working with sign language studies. However, we see that it is also being used by translators, interpreters, and general people who work in deaf education.

It is a platform designed to be integrated with sign language corpora, and it includes grammatical information associated with each sign of that database, with complex but easily manageable search tools. Considering the whole process, we also understood that planning for the sustainability of the platform is crucial. The plan is to share the signbank in its current state according to the same structure, and if one country decides to make feature improvements, these improvements ideally should apply to all countries using Signbank 2.0. This also makes it possible to create a network among all partners sharing Signbank 2.0. The Signbank network has two main innovative areas: the technological side and the linguistic side. The technology will be sustainably supported by a network, and the linguistic information shared between languages can feed sign language research worldwide.

### 4. Acknowledgements

Current research was funded partially by the National Brazilian Council for Scientific and Technological Development - CNPq (# 440337/2017-8) and the Department of Deaf Studies and Sign Language Interpreting, Humboldt-Universität zu Berlin. We acknowledge Onno Crasborn for developing Global Signbank that had been used as a base for developing Signbank 2.0 and the anonymous reviewers for the comments to improve this paper.

## 5. Bibliographical References

- Camargo, L. and Fazani, A. (2014). [Exploring the Participatory Design as a Support During the Development of Information Systems](#). In *InCID: Revista de Ciência da Informação e Documentação*, Ribeirão Preto, Brazil. 5(1):138–150.
- Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E. and Johnston, T. (2018). [Signbank: Software to Support Web Based Dictionaries of Sign Language](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2359–2364, Miyazaki, Japan. European Language Resources Association (ELRA).
- Crasborn, O., Hulsbosch, M. and Sloetjes, H. (2012). [Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS](#). In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 19–22, Istanbul, Turkey. European Language Resources Association (ELRA).
- Crasborn, O., Zwitserlood, I., Van Der Kooij, E. and Schuller, A. (2018). Global Signbank Manual. Version 1. Radboud University, Centre for Language Studies.
- Davidson, C. N. (2008). [Humanities 2.0: Promise, Perils, Predictions](#). In *Publications of the Modern Language Association of America*. 123(3):707–717.
- Fajardo, I., Parra, E. and Cañas, J. J. (2010). [Do sign language videos improve web navigation for deaf signer users?](#) In *Journal of Deaf Studies and Deaf Education*, 15(3):242–262.
- Flor, C. da S. (2016). Recomendações para a criação de pistas proximais de navegação em websites voltadas para surdos pré-linguísticos. In *Programa de Pós-graduação em Engenharia e Gestão do Conhecimento*, Universidade Federal de Santa Catarina, Florianópolis.
- Johnston, T. (1989). Auslan: the Sign Language of the Australian Deaf Community. Thesis Dissertation. University of Sydney, Sydney.
- Portal de Libras. (2021). <https://portal-libras.org>
- Quadros, R. M. de. (2016). Documentação da Libras. In *Anais Seminário Ibero-Americano de Diversidade Linguística. 2014*, Foz do Iguaçu. Brasília: IPHAN - Ministério da Cultura. 1:157–174.
- Quadros, R. M. de, Schmitt, D, Lohn, J. T. and Leite, T. de A. (2020). *Corpus de Libras*. <http://corpuslibras.ufsc.br/>
- Quadros, R. M. de, Krusser, D. and Saito, D. (2022). [Libras Portal: A Way of Documentation, a Way of Sharing](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 48–52, Marseille, France. European Language Resources Association,
- Rosenfeld, L., Morville, P., Arango, J. (2015). Information architecture: for the web and beyond. O'Reilly Media, 4th edition.
- Scolari, S. (2022). [O layout de configurações de mão em interfaces de busca](#). Tese (doutorado) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Design, Florianópolis.
- Scolari, S., Crasborn, O., Braviano, G. (2022). [Searching for Signs: Developing a handshape taxonomy based on visual similarity](#). In *International Journal of Lexicography*. 35(3):1–24.
- Wenger, E., McDermont, R., Snyder, W. M. (2002). Cultivating Communities of Practice: a guide to managing knowledge. Boston, Massachusetts: Harvard.



## Appendix: List of functionalities of Signbank 2.0

[for reference: <sup>1</sup> CRUD: Create, Read, Update, Delete ; <sup>2</sup> CRU: Create, Read, Update; <sup>3</sup> RUD: Read, Update, Delete; <sup>4</sup> Basic search: (1) Search text; (2) Newest order; (3) Order order; (4) Alphabetical order.]

The modules were listed in order of importance during the creation of the system. The planning was organized in terms of what would add value to each project with the following characteristics: (1) Project of sign language study; (2) Project of sign study for consultation; (3) Project adaptable to different sign languages; (4) Project adaptable to different institutions; (5) Project adaptable to different countries; (6) Project for the deaf people; (7) Open-source requirement; (8) Low resources for development maintenance.

### 1. Account

1.1. Register;

1.2. Login;

1.3. Forgot password;

1.4. Edit e-mail;

1.5. Edit password;

1.6. Confirmation password;

1.7. Delete account;

### 2. User

2.1. Edit some information user;

2.2. Change status account user;

2.3. Change type account user;

2.4. Change permission approver user;

2.5. Records who edit the user;

2.6. Read;

2.7. Search:

2.7.1. Basic search<sup>4</sup>

2.7.2. Pending Status;

2.7.3. Activated Status;

### 3. Modules Categories Provided to User:

3.1. Institution:

3.1.1. CRUD<sup>1</sup>;

3.1.2. Search:

3.1.2.1. Basic search<sup>4</sup>.

3.2. Profession:

3.2.1. CRUD<sup>1</sup>;

3.2.2. Search:

3.2.2.1. Basic search<sup>4</sup>.

3.3. Language:

3.3.1. CRUD<sup>1</sup>.

### 4. Signs:

4.1. CRUD<sup>1</sup> Signs

4.2. Search signs:

4.2.1. Handshape;

4.2.2. Basic search<sup>4</sup>;

4.2.3. Categories:

4.2.3.1. Handshape of dominant hand;

4.2.3.2. Handedness (searchable);

4.2.3.3. Location (area);

4.2.3.4. Movement shape;

4.2.3.5. Orientation change;

4.2.3.6. Relation between manual articulators;

4.2.3.7. Word class;

4.2.3.8. Semantic field;

4.2.3.9. Variation.

4.2.4. Pending Status;

4.2.5. Published Status;

### 5. Categories Signs:

5.1. Semantic Field:

5.1.1. CRUD<sup>1</sup>;

5.1.2. Search:

5.1.2.1. Basic search<sup>4</sup>.

- 5.2. Handshape:
  - 5.2.1. CRUD<sup>1</sup>;
  - 5.2.2. Search:
    - 5.2.2.1. Basic search<sup>4</sup>.
- 5.3. CRUD<sup>1</sup> Handedness:
  - 5.3.1. CRUD<sup>1</sup>;
  - 5.3.2. Search:
    - 5.3.2.1. Basic search<sup>4</sup>.
- 5.4. CRUD<sup>1</sup> Handedness (searchable group):
  - 5.4.1. CRUD<sup>1</sup>;
  - 5.4.2. Search:
    - 5.4.2.1. Basic search<sup>4</sup>.
- 5.5. CRUD<sup>1</sup> Handshape change:
  - 5.5.1. CRUD<sup>1</sup>;
  - 5.5.2. Search:
    - 5.5.2.1. Basic search<sup>4</sup>.
- 5.6. CRUD<sup>1</sup> Location Specific:
  - 5.6.1. CRUD<sup>1</sup>;
  - 5.6.2. Search:
    - 5.6.2.1. Basic search<sup>4</sup>.
- 5.7. CRUD<sup>1</sup> Location (area):
  - 5.7.1. CRUD<sup>1</sup>;
  - 5.7.2. Search:
    - 5.7.2.1. Basic search<sup>4</sup>.
- 5.8. CRUD<sup>1</sup> Relationship between manual articulators:
  - 5.8.1. CRUD<sup>1</sup>;
  - 5.8.2. Search:
    - 5.8.2.1. Basic search<sup>4</sup>.
- 5.9. CRUD<sup>1</sup> Orientation change:
  - 5.9.1. CRUD<sup>1</sup>;
  - 5.9.2. Search:
    - 5.9.2.1. Basic search<sup>4</sup>.
- 5.10. CRUD<sup>1</sup> Relative orientation: location:
  - 5.10.1. CRUD<sup>1</sup>;
  - 5.10.2. Search:
    - 5.10.2.1. Basic search<sup>4</sup>.
- 5.11. CRUD<sup>1</sup> Movement direction:
  - 5.11.1. CRUD<sup>1</sup>;
  - 5.11.2. Search:
    - 5.11.2.1. Basic search<sup>4</sup>.
- 5.12. CRUD<sup>1</sup> Movement shape:
  - 5.12.1. CRUD<sup>1</sup>;
  - 5.12.2. Search:
    - 5.12.2.1. Basic search<sup>4</sup>.
- 5.13. CRUD<sup>1</sup> Mouthing:
  - 5.13.1. CRUD<sup>1</sup>;
  - 5.13.2. Search:
    - 5.13.2.1. Basic search<sup>4</sup>.
- 5.14. CRUD<sup>1</sup> Mouth gestures:
  - 5.14.1. CRUD<sup>1</sup>;
  - 5.14.2. Search:
    - 5.14.2.1. Basic search<sup>4</sup>.
- 5.15. CRUD<sup>1</sup> Contact type:
  - 5.15.1. CRUD<sup>1</sup>;
  - 5.15.2. Search:
    - 5.15.2.1. Basic search<sup>4</sup>.
- 5.16. CRUD<sup>1</sup> Category of entity classifier:
  - 5.16.1. CRUD<sup>1</sup>;
  - 5.16.2. Search:
    - 5.16.2.1. Basic search<sup>4</sup>.
- 5.17. CRUD<sup>1</sup> Lexical types:
  - 5.17.1. CRUD<sup>1</sup>;
  - 5.17.2. Search:
    - 5.17.2.1. Basic search<sup>4</sup>.
- 5.18. CRUD<sup>1</sup> Variations:
  - 5.18.1. CRUD<sup>1</sup>;
  - 5.18.2. Search:
    - 5.18.2.1. Basic search<sup>4</sup>.
- 5.19. CRUD<sup>1</sup> Compounding:
  - 5.19.1. CRUD<sup>1</sup>;
  - 5.19.2. Search:
    - 5.19.2.1. Basic search<sup>4</sup>.

- 5.20. CRUD<sup>1</sup> Notes:
  - 5.20.1. CRUD<sup>1</sup>;
  - 5.20.2. Search:
    - 5.20.2.1. Basic search<sup>4</sup>.
- 5.21. CRUD<sup>1</sup> Word Class:
  - 5.21.1. CRUD<sup>1</sup>;
  - 5.21.2. Search:
    - 5.21.2.1. Basic search<sup>4</sup>.
- 5.22. CRUD<sup>1</sup> Tags:
  - 5.22.1. CRUD<sup>1</sup>;
  - 5.22.2. Search:
    - 5.22.2.1. Basic search<sup>4</sup>.

- 6.7. Frequently Asked Questions
  - 6.7.1. CRUD<sup>1</sup>

## 7. System Modules

- 7.1. Explanatory Videos Language Sign
  - 7.1.1. RUD<sup>3</sup>
  - 7.1.2. Search:
    - 7.1.2.1. Basic search<sup>4</sup>;
    - 7.1.2.2. Disabled status;
    - 7.1.2.3. Activated stats.
- 7.2. Platform identity
  - 7.2.1. Update

## 5. EAFS

- 5.1. CRUD<sup>1</sup> Videos EAF
- 5.2. Read file EAF for to extract the Occurrences
- 5.3. Read Occurrences

## 6. Institutional Modules

- 6.1. About Signbank
  - 6.1.1. Edit
- 6.2. Privacy Policy
  - 6.2.1. CRUD<sup>1</sup>
- 6.3. Terms of Use
  - 6.3.1. CRUD<sup>1</sup>
- 6.4. Platform Manual
  - 6.4.1. CRUD<sup>1</sup>
- 6.5. Annotation Manual
  - 6.5.1. CRUD<sup>1</sup>
- 6.6. Contact
  - 6.6.1. CRUD<sup>1</sup>
  - 6.6.2. Search:
    - 6.6.2.1. News order;
    - 6.6.2.2. Older order;
    - 6.6.2.3. Closed status;
    - 6.6.2.4. Waiting status;
    - 6.6.2.5. Text.

# STK LSF: A Motion Capture Dataset in LSF for SignToKids

Clément Reverdy, Sylvie Gibet, Thibaut Le Naour

IRISA, Université Bretagne Sud  
Vannes, France  
{clement.reverdy, sylvie.gibet}@irisa.fr

## Abstract

This article presents a new bilingual dataset in written French and French Sign Language (LSF), called *STK LSF*. This corpus is currently being produced as part of the SignToKids project. The aim of this corpus is to provide digital educational tools for deaf children, thereby facilitating the joint learning of LSF and written French. More broadly, it is intended to support future studies on the automatic processing of signed languages. To define this corpus, we focused on several grammatical phenomena typical to LSF, as well as in tales usually studied by hearing children in the second cycle in France. The corpus data represent approximately 1 hour of recording, carried out with a motion capture system (MoCap) offering a spatial precision of less than 1 mm and a temporal precision of 240 Hz. This high level of precision guarantees the quality of the data collected, which will be used both to build pedagogical scenarios in French and LSF, including signing avatar videos, and for automatic translation of text into LSF.

**Keywords:** French Sign Language, LSF, corpus, motion capture, grammatical utterances

## 1. Introduction

The aim of the *SignToKids* project is to build digital pedagogical tools for the joint learning of French sign language (LSF) and written French for deaf children. Of course, this work cannot cover all the educational needs to be made available to the Deaf, but it does provide an initial response to this very ambitious issue.

While there are a number of studies presenting the various aspects of French sign language (LSF) grammar (Cuxac, 2000; Millet, 2019), there is currently no educational book or digital application enabling schoolchildren to learn both the grammar of LSF and that of written French, which seems necessary for access to the various subjects taught in schools. Nor is there any specific bilingual LSF/French corpus built specifically for this purpose.

Through this project, it therefore proved necessary to define a specific bilingual corpus adapted to the demands of LSF and French teachers for deaf children in primary and secondary schools, corresponding to cycles II and III for hearing children. Our objectives are to: 1) make it easy to work out the grammatical structures common to and specific to each language; 2) help the child understand how to express the same concept in both languages; 3) correspond to the expectations of the cycle's curricula (<https://eduscol.education.fr/127/langue-des-signes-francaise>).

As part of our digital tools, rather than using videos in LSF, we have chosen to use virtual signing characters, which we call signing avatars. In addition, in order to obtain a high degree of precision in the movements produced, and to ensure the quality of the pedagogical exercises to be built, we opted for the use of motion capture data (Mo-

Cap). Furthermore, as the project is ambitious, both linguistically and technically, we decided to build the dataset in four phases, the first to adjust the corpus construction methodology and capture protocol, then to provide data with an increasing level of complexity.

This article describes the first two parts of the corpus, called LSF-STK1 corpus. It includes a set of phrases in written French and LSF, as well as the corresponding MoCap data.

## 2. Related work

Signed languages are visual and gestural languages. Consequently, the two main techniques for effectively capturing the sign language gestures are video and motion capture (MoCap). The corresponding two types of data do not entail the same costs or quality, either in terms of data acquisition or post-processing, and give rise to different analysis and processing possibilities.

### 2.1. Video corpora

Video recording devices (RGB or RGBD) are inexpensive, easy to set up and not very intrusive for the people being recorded. In addition, new tools (Kartynnik et al., 2019; Cao et al., 2017) have recently been developed to infer human postures (poses of the skeleton) and facial expressions from 2D images, making it possible to obtain gesture-type information. In addition, recent advances in computer vision (e.g. SMPL-X (Pavlakos et al., 2019)) make it possible to infer 3D meshes, blendshapes, skeletons and their animation parameters (joint orientation, blendshape coefficients, etc.) from human video recordings.





Figure 1: A few postures of our signing avatar.

Video data is often the basic material for linguistic analysis and automatic processing of sign languages (e.g. automatic recognition of signed sentences, automatic translation of a sign language into the corresponding written language) using machine learning algorithms (e.g. deep neural networks).

Among the large-scale projects that have emerged over the last decade, several initiatives are worth mentioning. However, as the subject of this paper is an LSF mocap dataset, only video corpora dealing with LSF or Belgian French sign language (LSFB) are listed here. For other SLs, the reader is invited to refer to the Sign Language Dataset Compendium (Kopf et al., 2022), which provides a list of most existing video corpora up to 2022 and their main characteristics.

There are two main corpora of LSF. The Mediapi-skel (Bull et al., 2020) is a ~27h corpus performed by more than 100 signers with a vocabulary size larger than 17k. It is signed by deaf journalists for a TV journal. The data are annotated with aligned written French subtitles. The CREAGEST (Balvet et al., 2010) is a corpus of ~500h signed by ~250 signers (adults and children), recorded in studio conditions, and elicited by various tasks. Within this corpus, only ~1h is annotated<sup>1</sup>.

The LSFB dataset (Fink et al., 2021) is a corpus in Belgian French sign language. It contains ~90h of videos, performed by 100 signers. ~25h data are recorded in studio conditions with a vocabulary of ~7k words. Elicitation was carried out by asking the signers to perform various tasks, leading to spontaneous discourse. These data are annotated with glosses and written French translations.

## 2.2. MoCap corpora

MoCap systems make it possible to record human gestures with a degree of precision, consistency and robustness not yet possible with video devices.

<sup>1</sup>according to <https://www.sign-lang.uni-hamburg.de/lr/compendium/corpus/creagest.html> dated January 23, 2023

Indeed, MoCap technologies have spatial accuracy in the millimeter range and frequency accuracy in the hundreds of Hertz range (typically 60 to 200 Hz for SL (Lefebvre-Albaret et al., 2013)). They are also less prone to occlusion problems than monovision devices. The limitations of this technology are characterized by: i) the need for data post-processing to reconstruct trajectories and produce skeletal pose sequences, which requires a considerable human investment (around one working day per minute of recording), and ii) the complexity of setting up this device, which limits its use to laboratory environments. This makes recordings less flexible than those obtained with lighter, less intrusive devices such as video. As a result, creating large corpora of MoCap data is still too costly for the time being, both in terms of equipment and human labour.

Among the MoCap corpora collected over the last decade, several have been produced with the aim of analyzing SL data and performing data-driven synthesis.

CUNY ASL (Lu and Huenerfauth, 2010, 2014) is an American Sign Language (ASL) dataset performed by 8 signers with a total duration of ~3h30. Elicitation was made by a native ASL speaker sitting behind the camera who engaged a conversation with the recorded signer. The body and fingers movements (using cybergloves ®), as well as gaze direction were tracked but facial expressions were not recorded. The data were annotated with glosses and spatial references.

(Jedlička et al., 2020) have recorded a ~30min full-body (body, face and fingers) MoCap dataset, performed by one expert Czech Sign Language (CSE) signer who was asked to sign weather forecasts. Data were annotated with glosses.

HRI JSL full-body dataset (Brock and Nakadai, 2018) has been performed by one signer. About 10k signed utterance in Japanese Sign Language (JSL) were recorded. The signer, a Child Of Deaf Adults (CODA) was asked to sign predefined sentences. Data was annotated with sign-based glosses.

In LSF, several corpora have been recorded. MO-CAP1 (Braffort, 2016) has been created for motion and linguistic analysis. Limbs motion and significant LSF facial movements were tracked, but fingers motion was not recorded. LSF-rosetta full-body corpus (Bertin-Lemée et al., 2022) aims to produce LSF from AZee specification (Nunnari et al., 2018), with a signing avatar. ~3h has been recorded by one signer. Elicitation was carried out by asking the signer to perform 4 tasks (translation to LSF, description of images, repetition of LSF video clips, production of >1200 isolated signs). Data was annotated with glosses, phonological components and AZee descriptions.

Several full-body high resolution LSF datasets have been recorded since 2009 at IRISA <sup>2</sup>, for LSF analysis and data-driven synthesis. For these datasets, skeletons were reconstructed from about 40 markers for the body, 40 smaller markers for the face, and 20 even smaller markers for each hand. Data-driven animations with a signing avatar were produced, after retargeting, rigging and skinning. LSF-SignCom (Duarte and Gibet, 2010) is a MoCap dataset of ~1.5h signed by one deaf signer. Based on a dialogue between two deaf signers, the movements of one of the two protagonists were recorded, with the second giving the cues. The corpus contains recipes and short stories on cooking themes (making salads, galettes, cocktails). Data was annotated on multiple tiers using glosses and phonological components. LSF-ANIMAL (Naert et al., 2020) is a full-body LSF dataset containing ~1h of data recorded on two deaf LSF professors fluent in written French. Elicitation was carried out with 3 main tasks: 1) isolated signs, 2) utterances illustrating grammatical mechanisms (pointing gestures, classifier predicates) and 3) Continuous signing (26 free descriptions of animals). Data was manually annotated with glosses on 3 channels (right hand, left hand and both). Phonological components (hand configuration, placement) were also annotated using automatic segmentation methods (Naert et al., 2018).

Several works have shown that data-based methods are capable of producing realistic LSF animations by concatenative synthesis. Thanks to a scripted language based on two coupled databases (motion and semantic annotations), new LSF utterances that respect LSF grammatical rules could be synthesized by editing and concatenating recorded motion, and used to animate a signing avatar (Gibet et al., 2011). Following the same approach, the *Sign3D* dataset has been recorded at Mocaplab. It contains utterances describing places of interest and events taking place in a city (Lefebvre-Albaret et al., 2013). Another recent project (Bertin-Lemée

---

<sup>2</sup>(Institut de Recherche en Informatique et Systèmes Aléatoires)

et al., 2022) follows a similar path.

The choice of MoCap therefore appears to be an appropriate solution for the recording of our corpus, both for the naturalness of the movements recorded, and for the precision of these movements, which meets the grammatical requirements of sign languages.

### 3. Corpus Design

The design of our corpus is based on the following approach: using a few a priori grammatical objectives, we construct a set of sentences which we then record (video, MoCap). But, rather than recording our corpus all at once, we have chosen to record four sub-corpora over a longer period - STK1.1, STK1.2, STK2.1, STK2.2 - in order to draw technical, pedagogical and linguistic lessons as we go along.

This paper refers to the first two sub-corpora that have been already recorded. Our first bilingual LSF/French sub-corpus, STK1.1, is based on several grammatical objectives that LSF teachers consider useful for learning both LSF and French. Of course, for each grammatical target, not all the processes used in French or LSF are covered exhaustively. For the creation of this sub-corpus, we constructed a set of sentences, guided by the use of teaching resources defined by deaf teachers in LSF (Centre Gabriel Deshayes, Auray) or from Millet's descriptive grammar (Millet, 2019). Our second sub-corpus, called STK1.2, describes three tales in LSF.

#### 3.1. Motivations for our Grammatical Targets in STK1.1

**Clausal aspects.** We first looked at the clausal form of sentences, i.e. negative, assertive and interrogative sentences.

The construction of negation is difficult for deaf children to learn, as it is expressed very differently in LSF and French. In French, negation is carried by the structures "ne ... pas", "ne ... plus", "ne ... jamais", etc. that surround the verb. In LSF, a distinction is made between negative sentences in which: 1) negation is marked by a specific lexical sign, generally placed at the end of the sentence, and those in which 2) negation is integrated into the verb. In addition, the expression of negation in LSF is generally accompanied by a facial mimic. The advantage of recording two logical forms of sentences in our data, positive and negative, enables us to construct simple exercises in which the negative form is requested from the positive form, in French or in LSF, and vice-versa.

Interrogative sentences are also used in exercises (for example question-and-answer exercises).

In this case, the question is direct, i.e. it is represented by a sign from the LSF lexicon. These signs can also be used in synthesis editing processes in rhetorical questions (false questions in LSF), which are not interrogative, but which serve to link propositions together. Here too, facial expressions are crucial, as they can be the unique mark of the interrogative clause as opposed to the assertive clause.

**Indicating verbs.** In STK1.1, we were interested in directional verbs that represent in French the syntactic structure Subject - Verb - Complement (direct or indirect object complements), with the possibility of representing subjects and recipients by pronouns. These verbs unfold along a trajectory in the signing space. This enables them to distribute syntactically the roles of the actants of the sentence (actants being the agent, the object, or the recipient). In this way, they move from one locus (spatial referent) to another. For example, the verb [TO GIVE] can be flexed along different trajectories from an agent locus to a recipient locus, and these actants can be pronouns positioned in the signing space. For example, the French sentence *Je te donne* (I give you) can be translated into LSF by a movement of the hand from a neutral zone near the torso (person 1) to a zone in front of the signer (person 2), while the sentence *Tu me donnes* is translated by an inverted trajectory. Some indicating verbs can also be flexed along the object. In this case, the configuration of the hand representing the object is modified. For example, the French sentences *Je te donne un verre* or *Je te donne un livre* are translated into LSF in the same way, except that the hand configuration changes to represent either the glass or the book.

The second STK1.2 sub-corpus is not associated with specific grammatical targets. It includes those of the STK1.1 corpus and other grammatical mechanisms typical to sign languages.

### 3.2. Corpus Content

**STK1.1** In the negation category, we have identified the signs [NON] (no), [JAMAIS] (never), [RIEN] (nothing), [Y-A-PAS] (nothing). Some signs, such as [NON], combine exclusively with verbs, and others with nouns, such as [Y-A-PAS]. For example, in the French sentences "Il ne boit pas" (He doesn't drink), "Elle ne boit jamais" (She never drinks), or "Je ne comprends rien" (I don't understand anything), negation is expressed in French by the words NE... PAS, NE... JAMAIS, NE... RIEN which surround the verb, whereas in LSF, this negation is expressed by a word at the end of the sentence ([NON], [JAMAIS], [Y-A-PAS]).

In the second negation category, we considered sentences with modalities, including:

[NE-PAS-POUVOIR] (can't), [NE-PAS-VOULOIR] (won't), [NE-PAS-CROIRE] (don't believe), [NE-PAS-SAVOIR] (don't know), [NE-PAS-AVOIR-BESOIN] (don't need). We have also added the verb [NE-PAS-AIMER] (dislike). Par exemple, dans la phrase "Le garçon n'aime pas facebook" (The boy doesn't like facebook), la négation s'exprime en LSF par [FACEBOOK][DISLIKE], dislike being represented by a sign, negative form of like. For example, in the sentence "Le garçon n'aime pas facebook" (The boy doesn't like facebook), the negation is expressed in LSF as [FACEBOOK][NE-PAS-AIMER], the negation being incorporated into the verb [NE-PAS-AIMER] (dislike), whose trajectory is reversed relatively to the verb [AIMER] (like). In another example, the sentence in French "Il n'a pas besoin qu'on lui dise deux fois" (He doesn't need to be told twice) can be translated in LSF as [REPETER][DEUX][FOIS][IL][NE-PAS-AVOIR-BESOIN], where the negation is directly incorporated into the verb [NE-PAS-AVOIR-BESOIN].

In both negation categories, we have defined 16 positive and 16 negative assertive sentences for each verb, and we have selected 24 interrogative sentences.

For indicating verbs, we selected 128 sentences repeated twice. For example, the French sentence "Je te raconte une histoire" (I am telling you a story), can be syntactically modified by replacing the pronouns "Je" (I) and "te" (you), as in "Elle me raconte une histoire" (She is telling me a story), or "Tu lui racontes une histoire" (You are telling her a story). In LSF, the syntactic structure with pronoun changes results in indicating verbs whose trajectories are modified, with personal pronouns (Je, Elle, Tu) or complement pronouns (te, me, lui) resulting in "pre-semanticized" Loci placed at specific places in the signing space. We have built a total of 128 different sentences with indicating verbs, declining the agent/object/beneficiary actants in each type of sentence. In the sub-corpus STK1.1, we have built a total of 220 phrases. All these sentences have been repeated twice in the recording session.

To this corpus we have added a list of words from a lexicon, which are chosen in such a way as to be able to construct new sentences in relation to the initial ones, thanks to the use of our concatenative synthesis system *SignCom* (Gibet et al., 2011; Naert et al., 2021).

Finally, we considered pointing to be fundamental to syntax, in particular to ensure reference. Pointing in LSF can be used in many semantic-syntactic contexts. In addition, there are many different ways of pointing. We have supplemented STK1.1 with a set of simple pointing gestures (index hand configuration) whose loci are randomly distributed in the signing space. About 40 pointing gestures have been executed.



**STK1.2** In order to have utterances more spontaneous (less controlled), with various grammatical structures proper to LSF (for example morpho-syntactic variations, iconic descriptions, static and dynamic classifiers), we produced STK1.2, composed of three tales that are usually studied in Cycle II for hearing children. The first two tales are French tales taken from the "Roman de Renart", a medieval collection of animal stories written by various authors. We have selected the two stories "Renard et la queue du loup" (Renard and the wolf's tail), and "Renard et les marchands" (Renard and the merchants). These tales were designed and adapted in French and in LSF, and validated by a deaf teacher in LSF. Both of them have a duration of about 10 minutes. Elicitation was then achieved by producing small written sequences (about 10 sentences per sequence) that tell the story, associated to glosses and illustrated by images. Sequences of questions are signed at the end of each tale.

The third tale follows another approach as it is directly signed in LSF by the deaf signer and then transcribed into written French. This tale is an adaptation of the tale "Le vilain petit canard" written by Hans Christian Andersen. Its duration is about 10 minutes. The three tales contain a total of 184 utterances.

## 4. Recording and Motion Dataset

Recording our corpus and motion dataset requires careful focus not only to the motion capture processing chain, but also to the development of tools for visualizing and editing motion, including the design of appropriate avatars.

### 4.1. Motion capture recording and processing

**Recording setup.** The *Motion-Up*<sup>3</sup> company was involved in the project to capture the data, model the 3D avatar and do the rigging and skinning. A total of one hour of data has been recorded over two sessions (STK1.1 and STK1.2). The MoCap system used for both recording sessions was an Optitrack 18-cameras "Prime 22" with reflexive markers, recording at 240 Hz.

To capture facial expressions, we explored two solutions: i) a MoCap-based solution, following the setup and approach described in (Reverdy, 2019), and ii) a commercial software (Faceshift (Weise et al., 2011)) that uses an RGBD video as input. Both options were tested during the initial session, and we chose to use the Faceshift solution exclusively for STK1.2, for reasons of simplicity and efficiency. Indeed, Faceshift provides, through a calibration process, a modeling of the human head with

51 morphotargets, and the automatic transformation of video into 51 blendshape curves.

In addition the scene was recorded by two RGB video cameras (60hz) from two different points of view in order to facilitate the annotation task. To make possible the synchronization between all devices, it was asked to the signer to perform at the beginning and the end of each recording sequence a specific mouth and hands movements.

**Markers setup.** 35 markers were placed all over the body, 40 on the signer's face (only during the first session) and 20 on each hand (a total of 75 markers on the body and hands). The number, size and shape of the markers used for each location is a trade-off between ease of tracking and comfort for the signer.

**Elicitation protocol.** The STK1 corpus was signed by a deaf signer who is also a theater actress. She participated in the design of the corpus content, in both LSF and French, and therefore had a good knowledge of it before each of the recording sessions. A slideshow of the sequences of sentences was projected at a distance of around 3 meters in front of the signer. For memorization purposes, she was instructed to sign the sentences presented in both written French and in the form of a sequence of glosses that she had previously transcribed and, where possible, illustrated with images.

The corpus was divided into sections of 1 to 4 slides, depending on the content of the corpus, with an average of 10 sentences per section. Each slide was composed of a set of sentences or isolated signs. Each section was repeated twice and recorded in one MoCap sequence. To achieve a certain quality of data, we made several takes for each section, but retained only one take for post-processing.

**Post-processing.** The data obtained after recording usually consists of a set of unlabeled marker trajectories. It is possible that, during recording, a marker is temporarily lost by the cameras for a short period (occultation), or that markers are exchanged along the trajectories.

Data cleaning involves labeling the markers along the trajectories and reconstructing the missing parts of the trajectories (gap-filling). In order to efficiently post-process the data, MotionUp has developed a software tool that reduces post-processing times by a 4x factor compared with the software *Motive 3.0.2 (Optitrack system)*. In particular, a new translation-and-rotation-invariant algorithm has been integrated into this Motion-Up software to automatically label the markers and reconstruct the trajectories. This very tedious task

---

<sup>3</sup><https://www.motion-up.com/>





Figure 2: An overview of the editing and visualization application.

required 0.6 working days per minute recorded for the body and hands.

Another tedious task was to clean up the facial data, as the mono-view recording device was prone to occlusions. These gaps had to be filled manually to ensure the quality of the animation. This task took an average of 0.72 days per recorded minute.

#### 4.2. Avatar design, Editing and Visualization Software

**Avatar Design.** The final objective of the *Sign-ToKids* project is to provide digital tools for learning both LSF and written French. From the point of view of educational exercise construction, we favor the inclusion of 3D virtual character animations as learning material, due to their interactive and playful aspect (possibility to modify the avatar's appearance) and the ease of editing this type of animation (e.g. camera movement, choice of reading speed).

The design of the avatar required careful technical consideration in order to limit the post-processing required for retargeting (adaptation to the signer's morphology) and to avoid animation artifacts that would damage the credibility and intelligibility of the signed gestures.

Morphologically, the joints and bone lengths have been finely designed to match the signer's skeleton. Furthermore, the face has been designed to be comparable to the signer's face, so that in the event of contact between hands and face, there is no gap or interpenetration.

**Editing and visualization software.** In addition to markers auto-labeling and gap-filing, Motion-Up software offers several other functionalities, including avatar visualization, editing and ultimate motion correction. An illustration is shown in Figure 2. This software features its own real-time iterative skeleton solver, so that the resulting animations can be played back, according to the selected SL sequence and the 3D avatar, and viewed in real time. The solver relies on an iterative process aimed at preserving the accuracy of spatial configurations such as contact between fingers, hands or face. Two mutually dependent steps were used, the first one for the palm transformation, the second for the fingers. This software also facilitates the editing of the recorded gesture afterwards, in order to correct certain undesirable behaviors (defaults generated by the MoCap installation, or signing defaults such as erroneous blinking or a systematic tendency to look in one direction).

### 5. Conclusion and Perspectives

This article presents an initial corpus STK1 and its corresponding dataset developed within the *Sign-ToKids* project. This corpus covers various grammatical phenomena essential to the joint learning of LSF and French. It also contains LSF adaptations of tales studied by hearing children aged 9 to 12. The chosen capture method is MoCap, due to the various objectives and constraints inherent in this project.

Approximately one hour of data was recorded

and post-processed, using a motion capture system, including body and hand motion, as well as facial expression and gaze direction. The skeleton data was reconstructed, and a 3D avatar was modeled and rigged to this data. An application was then developed to visualize, correct and annotate the various animations of this avatar corresponding to the recorded data. The linguistic annotation of this corpus has yet to be carried out with an annotation scheme inspired by the approach developed for the German Sign Language (DGS) corpus (Hanke et al., 2020).

The second part of this corpus, completing the recordings, is scheduled for the end of 2024. In the near future, we plan to expand this corpus by automatically producing, through generative AI, French and LSF-glossed sentences, and then by automatically translating the glossed sentences into LSF.

## 6. Acknowledgements

The *SignToKids* project is funded by the French National Research Agency (ANR). Several research teams participate in this project: IRISA-CNRS lab., MoDyCo-CNRS lab., and the national institute for young deaf (INJS). We would also like to thank Nathalie Irdel, from the Morbihan Deaf Association, for her help in designing and recording the *STK LSF* dataset.

## 7. Bibliographical References

- A. Balvet, C. Courtin, D. Boutet, C. Cuxac, I. Fusellier-Souza, B. Garcia, M.T. L'Huillier, and M.A. Sallandre. 2010. The creagest project: a digitized and annotated corpus for french sign language (lsf) and natural gestural languages.
- É. Bertin-Lemée, A. Braffort, C. Challant, C. Danet, B. Dauriac, M. Filhol, E. Martinod, and J. Segouat. 2022. Rosetta-lsf: an aligned corpus of french sign language and french for text-to-sign translation. In *13th Conference on Language Resources and Evaluation (LREC 2022)*.
- A. Braffort. 2016. Mocap1: A motion capture corpus of french sign language for interdisciplinary studies. In *Journées d'études DIG: La dynamique Interactionnelle du Geste "Making sense together"*.
- H. Brock and K. Nakadai. 2018. Deep jslc: A multimodal corpus collection for data-driven generation of japanese sign language expressions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- H. Bull, A. Braffort, and M. Gouiffès. 2020. Mediapi-skel-a 2d-skeleton video database of french sign language with aligned french subtitles. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6063–6068.
- Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. pages 7291–7299.
- C. Cuxac. 2000. *La langue des signes française (LSF) : les voies de l'iconocité (French) [French Sign Language: the iconicity ways]*. Faits de langues. Ophrys.
- K. Duarte and S. Gibet. 2010. Heterogeneous data sources for signed language analysis and synthesis: The signcom project. In *7th Int. Conf. on Language Resources and Evaluation (LREC 2010)*, volume 2, pages 1–8. European Language Resources Association.
- J. Fink, B. Frénay, L. Meurant, and A. Cleve. 2021. Lsfb-cont and lsfb-isol: Two new datasets for vision-based sign language recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- S. Gibet, N. Courty, K. Duarte, and T. Le Naour. 2011. The signcom system for data-driven animation of interactive virtual signers : Methodology and evaluation. In *ACM Transactions on Interactive Intelligent Systems*, volume 1.
- T. Hanke, M. Schulder, R. Konrad, and E. Jahn. 2020. Extending the public dgs corpus in size and depth. In *sign-lang@ LREC 2020*, pages 75–82. European Language Resources Association (ELRA).
- P. Jedlička, Z. Krňoul, J. Kanis, and M. Železný. 2020. Sign language motion capture dataset for data-driven synthesis. In *Workshop on the Representation and Processing of Sign Languages (LREC2020)*, pages 101–106.
- Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann. 2019. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*.
- M. Kopf, M. Schulder, and T. Hanke. 2022. The sign language dataset compendium: Creating an overview of digital linguistic resources. In *Workshop on the Representation and Processing of Sign Languages (LREC2022)*, pages 102–109.
- F. Lefebvre-Albaret, S. Gibet, A. Turki, L. Hamon, and R. Brun. 2013. Overview of the sign3d project high-fidelity 3d recording, indexing and editing of french sign language content. In *Third International Symposium on Sign*

*Language Translation and Avatar Technology (SLTAT) 2013.*

- P. Lu and M. Huenerfauth. 2010. Collecting a motion-capture corpus of american sign language for data-driven generation research. In *Proceedings of the NAACL HLT 2010 workshop on speech and language processing for assistive technologies*, pages 89–97.
- P. Lu and M. Huenerfauth. 2014. Collecting and evaluating the cuny asl corpus for research on american sign language animation. *Computer Speech & Language*, 28(3):812–831.
- A. Millet. 2019. *Grammaire descriptive de la langue des signes française: dynamiques iconiques et linguistique générale*. UGA Editions.
- L. Naert, C. Larboulette, and S. Gibet. 2020. Lsf-animal: A motion capture corpus in french sign language designed for the animation of signing avatars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6008–6017.
- L. Naert, C. Larboulette, and S. Gibet. 2021. [Motion synthesis and editing for the generation of new sign language content](#). *Machine Translation*, 35(3):405–430.
- L. Naert, C. Reverdy, C. Larboulette, and S. Gibet. 2018. Per channel automatic annotation of sign language motion capture data. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community (LREC 2018)*, Miyazaki, Japan.
- F. Nunnari, M. Filhol, and A. Heloir. 2018. Animating azee descriptions using off-the-shelf ik solvers. In *Workshop on the Representation and Processing of Sign Languages: Involving the Language Community (SignLang 2018)*, (LREC 2018), pages 7–12.
- G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A.A.A. Osman, D. Tzionas, and M.J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985.
- C. Reverdy. 2019. *Annotation et synthèse basée données des expressions faciales de la Langue des Signes Française*. Ph.D. thesis, Université de Bretagne Sud.
- T. Weise, S. Bouaziz, H. Li, and M. Pauly. 2011. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)*, 30(4):1–10.

# Preprocessing Mediapipe Keypoints with Keypoint Reconstruction and Anchors for Isolated Sign Language Recognition

Kyunggeun Roh<sup>1</sup>, Huije Lee<sup>1</sup>, Eui Jun Hwang<sup>1</sup>, Sukmin Cho<sup>1</sup>, Jong C. Park<sup>1</sup>

School of Computing

Korea Advanced Institute of Science and Technology

{rohbian, anquier, ehwa20, nellpic, jongpark}@kaist.ac.kr

## Abstract

Isolated Sign Language Recognition (ISLR) aims to classify signs into the corresponding gloss, but it remains challenging due to rapid movements and minute changes of hands. Pose-based approaches, recently gaining attention due to their robustness against the environment, are crucial against such challenging movements and changes due to the difficulty of capturing small joint movements from the noisy keypoints. In this work, we emphasize the importance of preprocessing keypoints to alleviate the risk of such errors. We employ normalization using anchor points to accurately track the relative motion of skeletal joints, focusing on hand movements. Additionally, we implement bilinear interpolation to reconstruct keypoints, particularly to retrieve missing information for hands that were not detected. Preprocessing methods proposed in this work show a 6.05% improvement in accuracy and achieved 83.26% accuracy with data augmentation on the WLASL dataset, which is the highest among pose-based approaches. The proposed methods show strengths in cases with signs having importance in the hand shape, especially when some frames have undetected hands.

**Keywords:** Sign Language Recognition, Keypoint Preprocessing, Transformer Architecture

## 1. Introduction

Sign language is the visual means of communication for the deaf, utilizing hand shapes, body movements, and facial expressions to convey messages. Like spoken languages, sign languages have their own diverse vocabulary and grammar. The difficulty of recognizing signs with detailed movements and diverse hand shapes remains as a barrier for hearing individuals to learn sign language. Sign Language Processing is an emerging field of machine learning that makes a bridge between the deaf and hearing individuals, by generating (Saunders et al., 2020), translating (Camgöz et al., 2020), and recognizing (Zhou et al., 2020) sign language expressions.

Isolated Sign Language Recognition (ISLR) focuses on translating sign language videos into the corresponding glosses, which are word-level representations of sign language expressions (Gobel and Assan, 1997; Jiang et al., 2021). ISLR shares similarities with video recognition tasks; however, the limited resources of ISLR datasets have been known as the main limitation, leading models to easily overfit on the dataset (Jang et al., 2022). Pose-based ISLR utilizes pose estimation models for keypoint extraction to overcome the challenges associated with the quantity and quality of datasets (Laines et al., 2023). The extracted keypoints remain independent of backgrounds and subjects, and since the keypoints are relatively lighter than RGB videos, they can also be easily augmented to prevent overfitting. Moreover, keypoints can be processed as sequential data with RNN or Transformer-

based models or as graph representations with graph neural networks (Ko et al., 2018; de Amorim et al., 2019).

Hand shape is one of the most important components of sign language, containing dense information in a smaller area than the body. Despite the importance of hand shape, pose-based approaches struggle with the challenging task of recognizing hand shapes, which easily differs with that of identifying minute movements of hand keypoints. To address this, the previous methods have been applying normalization on keypoints or have been implementing an additional model separately trained on the hands (Coster et al., 2020; Hu et al., 2021). The challenge becomes more difficult due to noisy keypoints from the failure of detection on the hands of the pose estimation model. For instance, Mediapipe (Lugaresi et al., 2019), a widely used pose estimation framework in the sign language domain, fails to detect over 50% of the hands appearing in each frame of the word-level American Sign Language dataset, WLASL. The noisy and undetected keypoints hinder hand shapes, leading to wrong predictions (Jiao et al., 2023).

In this work, we introduce a preprocessing framework, focused on hands, developed for pose-based ISLR. Our framework is based on the following strategies: anchor-based normalization, hand keypoint reconstruction, and fixing length. First, anchor-based normalization is applied to normalize the body and hands based on anchor points, which are set to clearly outline the hand shape by considering the relative distance between skeleton joints. Second, we employ keypoint reconstruc-



tion to recover the information of undetected hands by applying bilinear interpolation on surrounding frames. Additionally, the input sign language sequences are padded with frame duplication in a uniform distribution to train the model on stable data with a fixed length.

Finally, for evaluation, we validate our preprocessing framework on two representative ISLR datasets, WLASL-100 (Li et al., 2020a) and AUTSL (Sincan and Keles, 2020). The performance of the methods is assessed using both a Transformer encoder-decoder architecture and an encoder-only architecture to demonstrate the generality of the preprocessing methods. Our proposed methods improve the accuracy of recognizing sign language keypoints by 6.05%, and with basic augmentation, we achieve an accuracy of 83.26% on the WLASL-100 dataset, the highest among pose-based approaches. Further analysis demonstrates the significance of our normalization and reconstruction techniques in ISLR, and case studies show the effectiveness of our methods. We also discuss better input formats for sign language keypoints and handling highly undetected keypoints for future work.

## 2. Related Work

With the development of machine learning, ISLR research has also been highlighted in recent years. The approaches handling sign language videos are divided into two streams: the RGB-based approach, which directly recognizes features extracted from the RGB video into gloss representations, and the pose-based approach, which extracts skeleton keypoints from the RGB videos and recognizes the keypoints into the corresponding gloss.

### 2.1. RGB-based Approaches

Early Sign Language Recognition began with applying the Hidden Markov Model (HMM) to ISLR (Grobel and Assan, 1997). These approaches required additional equipment, such as colored gloves. However, with the development of CNN-based models, machine learning models can now segment the hand area without such additional equipment and directly extract feature vectors from the visual representation (Koller et al., 2018; Pigou et al., 2016). With the advancement of language processing models, the sequential feature vectors extracted from the CNN models can be effectively recognized with RNN or LSTM-based models (Koller et al., 2020; Cui et al., 2019). The development of 3D CNN models has demonstrated the strength of a single model capable of extracting both spatial and temporal information from videos without information loss between different models (Tran et al., 2015). Specif-

ically, research using the I3D model has shown that RGB-based methods can achieve reliable results in ISLR (Li et al., 2020a; Joze and Koller, 2019). Still, RGB-based approaches face limitations due to the constrained size of sign language video datasets. This leads models to develop biases towards the environments and appearances of the signers included in the training data. Recently, Jang et al. (2022) proposed a framework designed to augment the sign language video dataset by altering the background of the videos.

### 2.2. Pose-based Approaches

Pose-based ISLR has a significant advantage in that the pose estimation models are trained on a relatively large dataset compared to sign language datasets, making models more robust against different environments. Since the initial machine learning models with CNN architectures were not specifically designed to handle sequential keypoints, Pham et al. (2019) applied a transformation to the 3D skeleton keypoints to generate an image that contains both the spatial and temporal information of the keypoints, and a ResNet model was employed to recognize the generated image. With the enhancement of sequential models, the keypoints can be directly recognized with RNN or LSTM models (Ko et al., 2018; Liu et al., 2016; Papadimitriou et al., 2023). The skeleton keypoints can also be treated as graphs, and Graph Convolutional Networks (GCNs) have shown the strength of the architecture compared to the previous LSTM and RNN models (Maruyama et al., 2021). Especially, Jiang et al. (2021) have shown that pose-based architectures can outperform 3D CNN-based architectures with GCNs specialized for sign language. With the successful application of Transformer models to keypoints by Hu et al. (2021) and Boháček and Hružík (2022), recently, there has been an increasing focus among researchers on exploring the application of the Transformer model.

### 2.3. Preprocessing Pose Keypoints

One of the advantages of using skeleton keypoints is the lightweight nature compared to RGB videos, making preprocessing much easier. Normalization is a basic preprocessing method, and Transformer-based models have shown that the normalized keypoints can significantly improve the performance (Boháček and Hružík, 2022). With data augmentation, keypoint data can be augmented using basic approaches such as rotation and Gaussian noise to prevent the model from overfitting with limited data (Coster et al., 2020). Other approaches have shown that extracting additional features, such as movement of joints or bone information, can help

the recognition (Jiao et al., 2023). As shown in various studies, preprocessing methods enable models to learn effectively and overcome problems related to the limited amount of data.

The primary challenge with pose keypoints is that the pose estimation model can easily fail to detect the correct hand keypoints. To address such errors, researchers have been exploring better frameworks and attempting to combine different modalities (Zuo et al., 2023; Kanakanti et al., 2023). Masking keypoints is another preprocessing method aimed at reducing the risk from error keypoints and making the model more robust on such keypoints (Jiao et al., 2023; Hu et al., 2021). While current approaches focus on optimizing the use of the keypoints, there has not been as much exploration into recovering error keypoints.

In action recognition and other domains, several preprocessing approaches have been developed to improve the quality of noisy keypoints and reconstruct them using autoencoder models (Li et al., 2019; Wu et al., 2020; Zhou et al., 2021). However, these approaches face challenges when applied to sign language keypoints, particularly due to the frequent occurrences of undetected hands that such models cannot easily reconstruct. For instance, the Mediapipe framework, one of the major pose estimation frameworks for sign language, fails to detect almost 50% of the hands on the WLASL dataset. This high rate of undetection necessitates the adoption of alternative preprocessing techniques for hand pose reconstruction.

### 3. Methodology

The main goal of the proposed work is to concentrate on a more efficient method to normalize and reconstruct the hand keypoints, thereby facilitating the training of the model. In this section, we outline the designed experiments and provide details on how we handled and normalized the data.

#### 3.1. Anchor Based Normalization

Previous keypoint normalization techniques have been focusing on normalizing the keypoints based on the average position of the center of the body and rescaling lengths based on the shoulder length (Yoon et al., 2019). Especially, Coster et al. (2020) and Boháček and Hruz (2022) have normalized the keypoints with bounding boxes by aligning the keypoints in a segmented box region. Unlike such approaches, we envision that an anchor point could let the model learn better with a standard point. For this purpose, we normalized the keypoints by shifting them to position the neck (center of the body) fixed on the center  $(0, 0)$ . To normalize the length information, we divided all values by the length of

the neck instead of the shoulders because the neck seemed to be moving less than the shoulders for the ISLR task which has less facial expressions. By centering and scaling, we normalized the skeleton keypoints against the position of the signer so that the model becomes robust no matter how close the signer is to the camera or aligned in some direction. The equation below outlines the normalization process where  $x_k$  and  $y_k$  are the x and y coordinates, respectively, for each skeleton keypoint. The zeroth keypoint is designated as the neck ( $k = 0$ ), and the first keypoint ( $k = 1$ ) is identified as the center of the head. The normalization formula is given by:

$$(x'_k, y'_k) = \frac{(x_k, y_k) - (x_0, y_0)}{|(x_1, y_1) - (x_0, y_0)|} \quad (1)$$

We also conducted separate normalization for the hands, utilizing anchors positioned on the palm. In sign language recognition, the significance of the hand primarily stems from its shape and position. Since the position of the hand is already incorporated into the body keypoints with the wrist keypoint, our focus for the hands should be on shape information rather than position. To achieve this, we chose to normalize the hands separately from the body, akin to the approach taken by Boháček and Hruz (2022), to reduce the weight of positional information and emphasize shape information. However, to capture hand shapes more efficiently, we introduced anchors to the palm and shifted the hands based on these anchors to eliminate positional information. The size of the hands, containing information such as the relative distance from the body, is not separately normalized as length.

#### 3.2. Hand Keypoint Reconstruction

Sign language videos often include rapid hand movements, leading to blurry frames. Extracting keypoints from such blurry frames frequently results in failures in pose estimation. Additionally, signs involve occlusions due to overlapping hands, producing one of the most challenging cases to estimate accurately. To address these challenges, previous research has primarily focused on masking techniques to enhance the model's robustness against noisy keypoints (Hu et al., 2021; Jiao et al., 2023). While these approaches concentrate on making the model robust against noisy keypoints, Laines et al. (2023) have recovered positional information by placing undetected hand keypoints into the position of the palm.

Our approach focuses on recovering the basic information of the hand shape through keypoint reconstruction, as illustrated in Figure 1. We use bilinear interpolation to fill in the empty hand keypoints based on the surrounding skeleton keypoints. To apply bilinear interpolation to frames lacking keypoint data, we require at least one preceding and

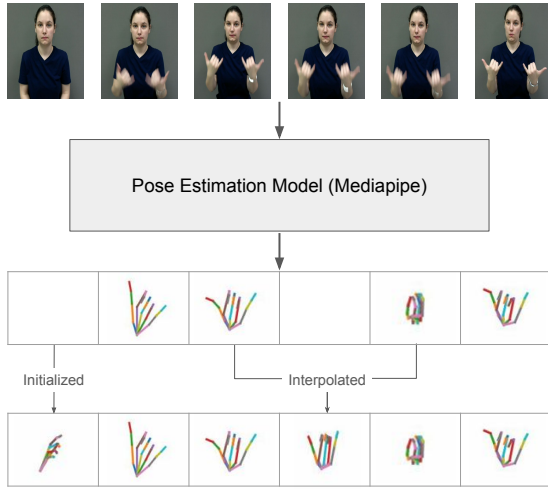


Figure 1: The process of initialization and reconstruction on a single hand. The average shape for the first and last frames is applied for initialization, and bilinear interpolation on other frames is used for reconstruction.

one succeeding frames with identified keypoints to serve as reference points. Therefore, we initiate our process by standardizing the keypoints of the first and last frames based on the average keypoint values, which typically represent the pose when the signer is waiting to start. This initialization step ensures that every empty frame is now sandwiched between frames populated with keypoints. Subsequently, we apply bilinear interpolation to these empty frames to recover the missing information. The provided equation for the normalized hand keypoints  $f_k$  from the  $k$ th frame incorporates a conditional mechanism to handle both the presence and absence of keypoint data. The equation is structured as follows:

$$f'_k = \begin{cases} \frac{\beta f_{k-\alpha} + \alpha f_{k+\beta}}{\alpha + \beta} & , \text{if } f_k = 0 \\ f_k & , \text{otherwise} \end{cases} \quad (2)$$

where  $\alpha$  and  $\beta$  are the minimum numbers that the  $k - \alpha$ th and  $k + \beta$ th frames have hand keypoints detected, respectively, which means  $f_{k-\alpha} \neq 0$ .

### 3.3. Fixing Length

One of the main motivations of this work is to concentrate on training the model more effectively through data preprocessing. We considered that methods related to the input length could also affect the model's performance. Sign language videos exhibit various lengths, ranging from below 15 frames to over 200 frames for a single gloss. The variability in length is due not only to the difficulty of expressing the sign but also to different signing styles among signers. Typically, padding is applied

Dataset	# Glosses	# Videos	Detect %
WLASL (2020a)	100	2k	46.56
AUTSL (2020)	226	36k	78.83

Table 1: Statistics related to the two datasets, WLASL and AUTSL. Detect % stands for the detection rate on hands, using the Mediapipe framework.

to short sequences to facilitate training together with long sequences in a single batch (Vázquez-Enríquez et al., 2021). Instead of padding, an alternative approach of interest was extending the length of the input sequence. To do so, frame duplication with a uniform distribution was applied to each instance, fixing the length to 512 frames.

## 4. Experiments

Experimental setups are introduced in this section. We provide information about the datasets used, the pose estimation frameworks employed for the experiments, and details regarding the settings.

### 4.1. Datasets

The datasets chosen to evaluate the proposed approaches are the WLASL and AUTSL datasets. WLASL is a Word-Level American Sign Language dataset that aligns with the task of ISLR (Li et al., 2020a). The dataset is structured with subsets of varying class sizes, 100, 300, 1,000, and 2,000 classes, which are ordered by the number of instances per class. Due to the difficulty of recognizing large subsets, which are unbalanced on the number of instances per class, we decided to use the smallest but richest subset, WLASL-100, for this experiment. The WLASL-100 dataset is composed of 2,038 instances from 97 different signers. With a relatively large number of signers, WLASL exhibits strength in diversity; however, this diversity makes recognition challenging due to the varying signing styles, speeds, and expressions.

The Ankara University Turkish Sign Language Dataset (AUTSL) is a Turkish Sign Language dataset with 226 classes, 36,302 instances, and 43 different signers (Sincan and Keles, 2020). The dataset is relatively balanced regarding the number of instances per class. However, with a smaller number of signers than WLASL, AUTSL exhibits limited diversity concerning the environment. These two datasets were selected for their distinct characteristics so that we can evaluate the efficacy of the proposed methods in diverse settings. As the datasets already include train/dev/test annotations, we apply the annotations for the experiment.

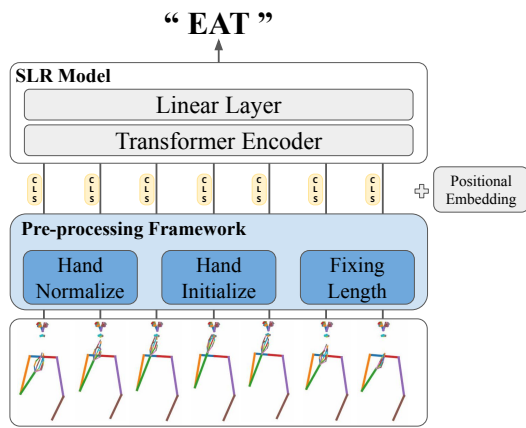


Figure 2: The baseline Transformer encoder architecture framework with preprocessing. Positional Embedding (PE) is added, and the CLS tokens are concatenated to the feature vectors.

## 4.2. Keypoint Representation

For the datasets mentioned in Section 4.1, we utilized Mediapipe Holistic to extract keypoints from the sign language videos (Lugaresi et al., 2019). Similar to the previous methods, we also decided not to use the z-axis data, as the Mediapipe documentation mentions that it is unreliable. The keypoints we used follow the work of Laines et al. (2023) for fair compatibility. As marked in Table 1, it is quite difficult to detect hands from sign language videos. To retain the positional information of the hands, the palm keypoints were duplicated to be included in the body, resulting in 20 keypoints for the face, 8 keypoints for the body (including the palms) and 21 keypoints for each hand, totaling 70 keypoints. For the baseline settings, the keypoints for undetected hands were set to the position of the palm, and for other settings, we preprocessed the keypoints as mentioned in Section 3.2.

## 4.3. Model Architecture and Setups

**Transformer Encoder.** As previous studies have demonstrated the effectiveness of applying Transformer architectures (Vaswani et al., 2017) to ISLR, we have also decided to utilize a Transformer architecture in this work (See Figure 2). However, our approach differs in that we only applied the encoder model, which appeared to be more efficient. The vanilla Transformer encoder model with 4 layers was applied, which shows reliable performance with low complexity that seemed to be more efficient than using more layers. Positional embedding was incorporated with learnable parameters to train the model with the awareness of spatial information, which indicates that each skeleton joint contains distinct information. Similar to the classification

based Transformer models by Devlin et al. (2019) and Dosovitskiy et al. (2021), class tokens are concatenated to the features as a parameter. Finally, a linear layer is applied to the output class tokens, and accuracy is measured. We set the Transformer encoder architecture as the baseline and demonstrate the effects of the proposed methods.

To compare the Transformer encoder model with previous researches based on a different Transformer model, we also employ the architecture of SPOTER (Boháček and Hružík, 2022). SPOTER is based on a Transformer encoder-decoder architecture with 6 layers and positional embeddings on every feature that contain both spatial and temporal information.

**Training Details.** The learning rate was fixed at  $1e-5$ , and the models were trained for 200 epochs. Batch size differed between datasets, with WLASL-100 trained on batch size 4, while AUTSL, which has a relatively larger size, was trained with batch size 16. Adam Optimizer was used for optimization. Cross-entropy loss was employed for the training loss, and the top-1 accuracy score was measured for evaluation. All results shared in the results are the average scores from 5 or more attempts with a random seed, as the results may vary depending on the seed number.

## 4.4. Data Augmentation

Data augmentation is considered as one of the distinct strengths of pose-based ISLR (Alyami et al., 2024; Selvaraj et al., 2022). Previous research has consistently demonstrated that data augmentation significantly enhances performance, especially on limited datasets with unbalanced instances (Zuo et al., 2023). By implementing data augmentation, we show that the proposed preprocessing methods are independent of data augmentation, which means that the methods can be utilized together with different data augmentation techniques from previous and future works.

In this study, we implemented widely adopted augmentation techniques, rotation and Gaussian noise. We adopted the augmentation settings as utilized by Boháček and Hružík (2022), applying rotation with angles randomly chosen between  $-13$  and  $13$  degrees and adding Gaussian noise to each keypoint, following a distribution with a mean of 0 and a standard deviation of  $10^{-3}$ .

# 5. Results and Analysis

## 5.1. Main Results

The results of applying the proposed methods appear in Table 2 on the two datasets. With the encoder-only model that we proposed, we can see



Dataset	Model	Method			Acc. (%)
		Hand Normalize	Hand Initialize	Fixing Length	
WLASL	Transformer Encoder-Decoder (SPOTER)	✗	✗	✗	71.63
		✓	✗	✗	79.38
		✓	✓	✗	<b>80.31</b>
		✓	✗	✓	78.68
		✓	✓	✓	79.46
	Transformer Encoder-only (Baseline)	✗	✗	✗	76.12
		✓	✗	✗	79.85
		✓	✓	✗	80.62
		✓	✗	✓	81.16
		✓	✓	✓	<b>82.17</b>
AUTSL	Transformer Encoder-only (Baseline)	✗	✗	✗	90.40
		✓	✗	✗	90.76
		✓	✓	✗	90.77
		✓	✗	✓	<u>90.95</u>
		✓	✓	✓	<b>91.15</b>

Table 2: Comparative results on WLASL and AUTSL between SPOTER and our Transformer encoder ISLR model under three preprocessing settings. Results with the best accuracy score are bold, and the following best results are underlined.

that normalizing hands based on anchors significantly improves accuracy with a 3.73% improvement on the WLASL dataset. Moreover, initializing the keypoints with bilinear interpolation and fixing the input length has also enhanced the performance. Applying all of the methods together, the encoder-only model has shown a 6.05% improvement. The performance change is relatively small in the AUTSL dataset; however, we notice that each method is improving the performance and showing a similar tendency with the results of WLASL.

Results from the Transformer encoder-decoder model show that anchor-based normalization and reconstruction of hands give rise to a significant improvement, which shows the generality of the two methods on a different model architecture. Unlike other methods, fixing the length seemed to be bothering the training process on the encoder-decoder model. The difference of the model based on the SPOTER architecture and the encoder-only model is that the positional embedding of the SPOTER has considered both the spatial and temporal embeddings together, while the baseline model has only been focusing on embedding spatial information. As the length of the input sequences has been extended and fixed by duplication, it seemed that the inconsistent information with the temporal embedding resulted in a lower performance.

## 5.2. Comparison with Other Methods

**WLASL.** Results conducted on WLASL are presented in Table 3. Our proposed method outperforms previous pose-based methods. SPOTER<sup>†</sup>

Method	Modality	Acc. (%)
I3D (2020a)		65.89
TK-3DConvNet (2020b)	RGB	77.55
Full Transformer (2022)		80.72
GCNBERT (2021)		60.15
SPOTER (2022)		63.18
SPOTER <sup>†</sup>	Pose	71.63
SignBERT (2021)		79.07
SL-TSSI <sup>†</sup> (2023)		81.47
I3D+ST-GCN (2021)		81.38
SignBERT (2021)	Multi.	82.56
NLA-SLR (2023)		<b>92.64</b>
Ours <sup>†</sup>	Pose	82.17
Ours <sup>†</sup> w/ Augment		<u>83.26</u>

Table 3: Accuracy comparison on WLASL with previous methods using different modalities. Note that the dagger(†) mark refers to researches based on Mediapipe keypoints and Multi. refers to the multimodal approaches.

is the result of the SPOTER model trained on Mediapipe keypoints. Our approach outperforms previous RGB-based methods and most of the multimodal methods that use pose and RGB data together. While we still cannot reach the performance of the NLA-SLR model by Zuo et al. (2023), the results highlight the importance of the proposed preprocessing methods.

**AUTSL.** In contrast to the results related to WLASL, the results presented in Table 4 indicate

Method	Modality	Acc. (%)
VTN-PF (2021)		92.92
I3D (2022)	RGB	93.53
MViT-SLR (2023)		95.72
SL-TSSI <sup>†</sup> (2023)		93.13
MS-G3D (2021)	Pose	95.38
SL-GCN (2021)		96.47
SAM-SLR (2021)	Multi.	<b>98.53</b>
Ours <sup>†</sup>	Pose	91.15
Ours <sup>†</sup> w/ Augment		91.66

Table 4: Accuracy comparison on AUTSL-100 with previous methods with different modalities. Note that the dagger(<sup>†</sup>) mark refers to researches based on Mediapipe keypoints and Multi. refers to the multimodal approaches.

Method	None	Gauss.	Rotate	Both
Accuracy	82.17	82.24	82.63	<b>83.26</b>

Table 5: Accuracy score with different data augmentation methods, Gaussian noise, rotation, and applying both.

that our model underperforms compared to previous methods based on pose and RGB data. The limitation seemed to be due to the smaller number of parameters than the previous methods, as it is highlighted in the earlier work of Laines et al. (2023). SL-TSSI employs 7.2M parameters, and SL-GCN employs around 19.2M parameters, whereas our method works on 5.3M parameters. Moreover, the difference based on the pose estimation frameworks shows that only SL-TSSI has been using the Mediapipe keypoints, which produces a relatively similar result compared to others.

### 5.3. Analysis

**Data Augmentation.** We also show that our methods can be enhanced with basic data augmentation skills mentioned in Section 3.4. Table 5 shares the results of applying each augmentation skill. Both augmentation methods are showing improvements, especially when they are applied simultaneously. These results demonstrate that the proposed methods and data augmentation complement each other and show the possibilities with more complicated augmentation methods, such as augmentation based on speed or joint rotation (Boháček and Hružík, 2022; Laines et al., 2023).

**Normalization Comparison.** To show the importance of the anchor-based normalization, we share the results of normalizing our model and the

Method	SPOTER	TF Encoder
Bounding Box	76.59	78.06
Anchor-based	<b>79.38</b>	<b>79.85</b>

Table 6: Accuracy score of the two models, SPOTER and our Transformer encoder model, with the two different normalization methods of setting bounding boxes and normalizing based on anchors.

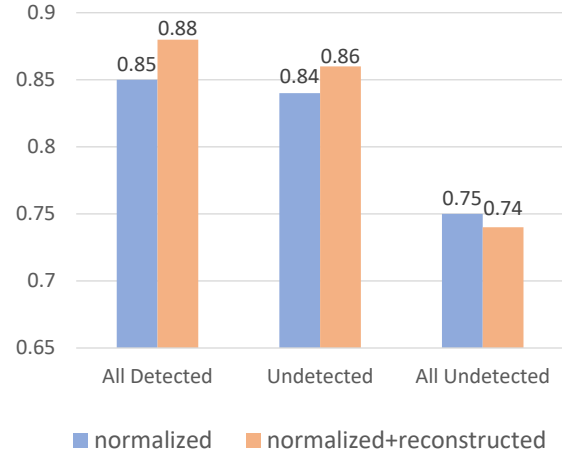


Figure 3: Accuracy scores on all detected, undetected, and all undetected cases. All detected stands for the instances that have all of the hands detected, undetected for those with some hands undetected, and all undetected for those that have at least one hand undetected in all of the frames.

SPOTER model based on our anchor-based normalization and the bounding box-based normalization in Table 6. As we can see, the normalization with anchors on the hands shows better performance on the two different models. The use of anchor keypoints suggests that the model learns more effectively based on the relative distance between skeleton joints.

**Reconstruction Effectiveness.** The proposed methods have shown improvements in the model performance. To clearly see that the model is recovering the information of keypoints, we divided the WLASL test dataset according to whether the hand detection fails or not. Instances with all hands well detected are checked as “all detected”, some frames having undetected hands are checked as “undetected”, and those with all frames having at least one hand undetected are checked as “all undetected”. For comparison, we analyzed our proposed methods trained with the hands normalized and having the hands reconstructed.

Results are shared in Figure 3, where we observe that the model trained on reconstructed hands exhibits the strength in instances where at least some

hands are detected. The reconstruction not only seemed to be improving the performance based on the recovered information but also seemed to be alleviating the difficulty of training the model with different keypoint representations, some of which have all keypoints detected while others are missing many of the keypoints. However, instances with almost no hands detected seemed to be struggling with reconstructed keypoints that do not possess much information, resulting in a slight decrease in performance. Still, the trade-off is smaller than the improvements noticing that the keypoint reconstruction recovers some information and alleviates the problems coming from undetected hands.

Case 1	Input Sequence						Gloss
Original							Pull
Extracted							Bowling
Ours							Pull
Case 2	Input Sequence						Gloss
Original							Graduate
Extracted							Help
Anchor-base Normalized							Graduate

Figure 4: Case studies on the WLASL dataset. Hand keypoints successfully reconstructed are highlighted with red boxes.

#### 5.4. Case Studies

Finally, case studies were conducted to determine if the proposed methods were successfully applied to specific cases. Figure 4 illustrates two cases of when our method has been applied successfully. The first case contains an example where some of the hands are undetected, leading to incorrect predictions. Empty hand keypoints confuse the model, causing it to predict the input sequence into glosses having similar body movements but different hand shapes. Pull and bowling serve as examples of such difficult cases with similar body movements. The loss of keypoints seemed to be leading the model to incorrect predictions. Keypoint reconstruction applied in the proposed research reconstructs the missing hand keypoints and leads the model to correct predictions.

The second case contains an example with a sign that has a particular hand shape containing

some important information while the body does not move so much. When the hands are not separately normalized based on anchors, the model struggles to predict similar signs having similar motions even though all hand keypoints are detected. Anchor-based normalization seemed to help the model recognize the shape of hands, leading to correct predictions.

## 6. Conclusions and Limitations

In this work, we proposed preprocessing methods for Isolated Sign Language Recognition (ISLR). First, we have applied anchor-based normalization, which normalizes the body and hands based on anchor points. Particularly, anchors from the hands remove unnecessary positional information and emphasize the distance between keypoints that effectively retains the shape information. Second, undetected hand keypoints were reconstructed using bilinear interpolation, showing that the reconstructed keypoints recover the shape information of hands. Finally, the length of the sign language sequence was fixed to relieve the difficulty of training a model on data with diverse input lengths. We argue that the methods show the generality across different model architectures and datasets. The application of basic data augmentation methods has improved the performance, demonstrating that the preprocessing methods are independent of data augmentation.

Still, we have several tasks to explore in the future. Fixing the length of the input sequence has been interrupting the training process when we applied the Transformer encoder-decoder model which has both spatial and temporal embeddings. We assume that the temporal embeddings have inconsistent information with the duplicated frames, and leave the question of implementing a better format instead of duplicating the frames for stable training on diverse models for future work. Additionally, the proposed methods still face challenges in cases with highly undetected keypoints, which need to be addressed as well in subsequent work by applying other preprocessing methods or better pose estimation frameworks specialized on hands (Ivashechkin et al., 2023).

## 7. Acknowledgements

This work was supported by the Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00010, Development of Korean sign language translation service technology for the deaf in medical environment).

## 8. Bibliographical References

- Sarah Alyami, Hamzah Luqman, and Mohammad Hammoudeh. 2024. [Isolated Arabic Sign Language recognition using a transformer-based model and landmark keypoints](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(1).
- Matyás Boháček and Marek Hruží. 2022. [Sign pose-based transformer for word-level sign language recognition](#). In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 182–191. IEEE.
- Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. [Sign language transformers: Joint end-to-end sign language recognition and translation](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.
- Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2020. [Sign language recognition with transformer networks](#). In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6018–6024. European Language Resources Association.
- Mathieu De Coster, Mieke Van Herreweghe, and Joni Dambre. 2021. [Isolated sign recognition from RGB video using pose flow and self-attention](#). In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3441–3450. Computer Vision Foundation / IEEE.
- Runpeng Cui, Hu Liu, and Changshui Zhang. 2019. [A deep neural framework for continuous sign language recognition by iterative training](#). *IEEE Trans. Multim.*, 21(7):1880–1891.
- Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri N. Metaxas. 2022. [Bidirectional skeleton-based isolated sign recognition using graph convolutional networks](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 7328–7338. European Language Resources Association.
- Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. 2019. [Spatial-temporal graph convolutional networks for sign language recognition](#). In *Artificial Neural Networks and Machine Learning - ICANN 2019 - 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17-19, 2019, Proceedings - Workshop and Special Sessions*, volume 11731 of *Lecture Notes in Computer Science*, pages 646–657. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An image is worth 16x16 words: Transformers for image recognition at scale](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Yao Du, Pan Xie, Mingye Wang, Xiaohui Hu, Zheng Zhao, and Jiaqi Liu. 2022. [Full transformer network with masking future for word-level sign language recognition](#). *Neurocomputing*, 500:115–123.
- Yong Du, Wei Wang, and Liang Wang. 2015. [Hierarchical recurrent neural network for skeleton based action recognition](#). In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1110–1118. IEEE Computer Society.
- Kirsti Grobel and Marcell Assan. 1997. Isolated sign language recognition using hidden markov models. In *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*, volume 1, pages 162–167. IEEE.
- Al Amin Hosain, Panneer Selvam Santhalingam, Parth H. Pathak, Huzefa Rangwala, and Jana Kosecká. 2021. [Hand pose guided 3D pooling for word-level sign language recognition](#). In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 3428–3438. IEEE.
- Hezhen Hu, Weichao Zhao, Wengang Zhou, Yuechen Wang, and Houqiang Li. 2021. [SignBERT: Pre-training of hand-model-aware representation for sign language recognition](#). In



- 2021 *IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11067–11076. IEEE.
- Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. [Denoising diffusion for 3d hand pose estimation from images](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3128–3137. IEEE.
- Youngjoon Jang, Youngtaek Oh, Jae-Won Cho, Dong-Jin Kim, Joon Son Chung, and In So Kweon. 2022. [Signing outside the studio: Benchmarking background robustness for continuous sign language recognition](#). page 322.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. [Skeleton aware multi-modal sign language recognition](#). In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3413–3423. Computer Vision Foundation / IEEE.
- Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, and Xilin Chen. 2023. [CoSign: Exploring co-occurrence signals in skeleton-based continuous sign language recognition](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20619–20629. IEEE.
- Cristina Luna Jiménez, Manuel Gil-Martín, Ricardo Kleinlein, Rubén San Segundo, and Fernando Fernández Martínez. 2023. [Interpreting sign language recognition using transformers and Mediapipe landmarks](#). In *Proceedings of the 25th International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9-13, 2023*, pages 373–377. ACM.
- Hamid Reza Vaezi Joze and Oscar Koller. 2019. [MS-ASL: A large-scale data set and benchmark for understanding American Sign Language](#). page 100.
- Mounika Kanakanti, Shantanu Singh, and Manish Shrivastava. 2023. [MultiFacet: A multi-tasking framework for speech-to-sign language generation](#). In *International Conference on Multimodal Interaction, ICMI 2023, Companion Volume, Paris, France, October 9-13, 2023*, pages 205–213. ACM.
- Sang-Ki Ko, Jae Gi Son, and Hye Dong Jung. 2018. [Sign language recognition with recurrent neural network using human keypoint detection](#). In *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems, RACS 2018, Honolulu, HI, USA, October 09-12, 2018*, pages 326–328. ACM.
- Oscar Koller, Necati Cihan Camgöz, Hermann Ney, and Richard Bowden. 2020. [Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2306–2320.
- Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. [Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn-hmms](#). *Int. J. Comput. Vis.*, 126(12):1311–1325.
- David Laines, Miguel González-Mendoza, Gilberto Ochoa-Ruiz, and Gissella Bejarano. 2023. [Isolated sign language recognition based on tree structure skeleton images](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 276–284. IEEE.
- Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020a. [Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison](#). In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pages 1448–1458. IEEE.
- Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersen, and Hongdong Li. 2020b. [Transferring cross-domain knowledge for video sign language recognition](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6204–6213. Computer Vision Foundation / IEEE.
- Shujie Li, Yang Zhou, Haisheng Zhu, Wenjun Xie, Yang Zhao, and Xiaoping Liu. 2019. [Bidirectional recurrent autoencoder for 3d skeleton motion data refinement](#). *Comput. Graph.*, 81:92–103.
- Tao Liu, Wengang Zhou, and Houqiang Li. 2016. [Sign language recognition with long short-term memory](#). In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 2871–2875. IEEE.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [Mediapipe: A framework for building perception pipelines](#). *CoRR*, abs/1906.08172.

- Mizuki Maruyama, Shuvozit Ghose, Katsufumi Inoue, Partha Pratim Roy, Masakazu Iwamura, and Michifumi Yoshioka. 2021. [Word-level sign language recognition with multi-stream neural networks focusing on local regions](#). *CoRR*, abs/2106.15989.
- Gokul NC, Manideep Ladi, Sumit Negi, Prem Selvaraj, Pratyush Kumar, and Mitesh M. Khapra. 2022. [Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Maxim Novopoltsev, Leonid Verkhovtsev, Ruslan Murtazin, Dmitriy Milevich, and Luliia Zemtsova. 2023. [Fine-tuning of sign language recognition models: a technical report](#). *CoRR*, abs/2302.07693.
- Katerina Papadimitriou, Gerasimos Potamianos, Galini Sapountzaki, Theodoros Goulas, Eleni Efthimiou, Stavroula-Evita Fotinea, and Petros Maragos. 2023. [Greek Sign Language recognition for an education platform](#). *Universal Access in the Information Society*, pages 1–18.
- Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin. 2019. [Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks](#). *IET Comput. Vis.*, 13(3):319–328.
- Lionel Pigou, Mieke Van Herreweghe, and Joni Dambre. 2016. Sign classification in sign language corpora with deep neural networks. In *sign-lang@LREC 2016*, pages 175–178. European Language Resources Association (ELRA).
- Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020. [Progressive transformers for end-to-end sign language production](#). In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 687–705. Springer.
- Prem Selvaraj, Gokul Nc, Pratyush Kumar, and Mitesh Khapra. 2022. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2133.
- Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2020. [AUTSL: A large scale multi-modal turkish sign language dataset and baseline methods](#). *IEEE Access*, 8:181340–181355.
- Ozge Mercanoglu Sincan and Hacer Yalim Keles. 2022. [Using motion history images with 3D convolutional networks in isolated sign language recognition](#). *IEEE Access*, 10:18608–18618.
- Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. [Learning spatiotemporal features with 3D convolutional networks](#). In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 4489–4497. IEEE Computer Society.
- Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan P. Wachs. 2021. [Pose-based sign language recognition using GCN and BERT](#). In *IEEE Winter Conference on Applications of Computer Vision Workshops, WACV Workshops 2021, Waikoloa, HI, USA, January 5-9, 2021*, pages 31–40. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Manuel Vázquez-Enríquez, José Luis Alba-Castro, Laura Docío Fernández, and Eduardo Rodríguez Banga. 2021. [Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks](#). In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 3462–3471. Computer Vision Foundation / IEEE.
- Zhize Wu, Thomas Weise, Le Zou, Fei Sun, and Ming Tan. 2020. [Skeleton based action recognition using a stacked denoising autoencoder with constraints of privileged information](#). *CoRR*, abs/2003.05684.
- Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. [Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots](#). In *International Conference on Robotics and Automation, ICRA 2019, Montreal, QC, Canada, May 20-24, 2019*, pages 4303–4309. IEEE.
- Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2020. [Spatial-temporal multi-cue network for continuous sign language recognition](#).

In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13009–13016. AAAI Press.

Kanglei Zhou, Zhiyuan Cheng, Hubert P. H. Shum, Frederick W. B. Li, and Xiaohui Liang. 2021. [STGAE: spatial-temporal graph auto-encoder for hand motion denoising](#). In *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2021, Bari, Italy, October 4-8, 2021*, pages 41–49. IEEE.

Ronglai Zuo, Fangyun Wei, and Brian Mak. 2023. [Natural language-assisted sign language recognition](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14890–14900. IEEE.

# Decoding Sign Languages: The SL-FE Framework for Phonological Analysis and Automated Annotation

Karahan Şahin, Kadir Gökğöz

Bogazici University

Bogazici University. Department of Linguistics. 34342 Bebek/İstanbul. Turkey

{karahan.sahin, kadir.gokgoz}@boun.edu.tr

## Abstract

SL-FE is a framework designed for the phonological representation of sign languages, bridging the gap between theoretical phonology and practical sign language annotation. SL-FE defines phonological information as a continuous signal from pose estimation information that enables not only the extraction of the comprehensive set of discrete phonological information but also provides a quantitative framework for theoretical analyses. By utilizing our framework, we conduct case studies to test empirical claims of feature dominance and symmetry on phonological complexity in Turkish Sign Language (TID). Only by defining a ranking function, we were able to classify these conditions with high lexical retrieval accuracy offering empirical evidence to support theoretical claims. The framework proves to be an essential tool for research in sign language linguistics.

**Keywords:** sign language phonology, automatic annotation, pose estimation

## 1. Introduction

The field of sign language research has seen considerable advancements in automatic annotation technologies, significantly enhancing the efficiency and accuracy of sign language recognition and translation. However, a gap persists in integrating theoretical phonological models into these frameworks. Traditional automatic annotation systems primarily focus on feature extraction, serving the immediate needs of recognition and translation without delving into the theoretical aspects of sign languages (Skobov and Lepage, 2020; Lucie Naert and Gibet, 2018; Gonzalez et al., 2012). While functional for specific applications, this approach overlooks the phonological information crucial for comprehensive linguistic analysis and understanding, with the .

In response to this need, our framework, **Sign Language Feature Extraction (SL-FE)**, emerges as a novel solution for the limitations of existing annotation systems. Unlike its predecessors, SL-FE is not merely an automatic annotation tool but a robust framework incorporating a continuous mathematical representation of phonological information specifically tailored for sign languages. Drawing upon prosodic models (Fenlon et al., 2017), SL-FE represents each phonological feature — finger selection, movement, and location information— through normalized feature-scoring methods. This method leverages pose-estimation technology to calculate the probability of feature occurrences, utilizing both orthogonal and angular distances between joints and normalizing these measurements according to the body proportions of the signer. Such an approach ensures that our scoring remains

invariant to variations in signer and camera angles, providing a consistent and interpretable analysis of phonological features in continuous sign language videos as demonstrated in Figure 1.

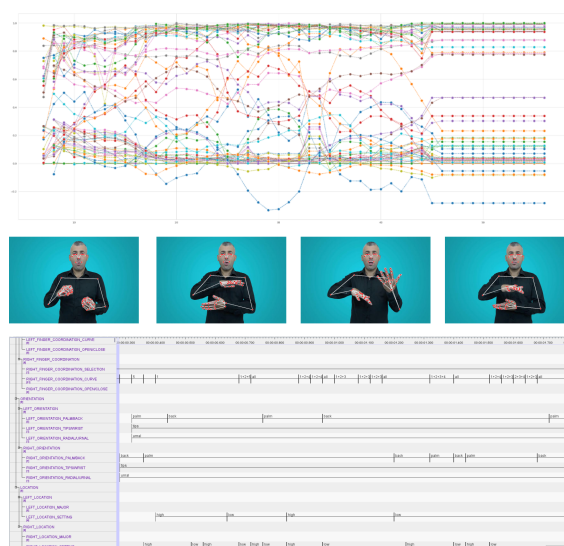


Figure 1: The pipeline of our framework for the lexical item "EVENT", in TID Sözlük. **The top side** is the cumulative plot of extracted continuous phonological information from the sign language video. **On the bottom side**, the annotations are exported to the ELAN interface after the classification pipeline is applied to the continuous feature set.

A significant achievement of our framework is its capacity to operationalize and validate typological claims within sign language research, such as Feature Dominance and Feature Symmetry (Bat-



tison, 1978). By applying SL-FE to the TID (Turkish Sign Language) Sözlük Dictionary database, we have successfully computed the phonological complexity of isolated lexical items, offering empirical support for these theoretical constructs. This capability not only demonstrates the framework's analytical power but also contributes to the broader understanding of sign language phonology.

Furthermore, SL-FE is designed with accessibility in mind. The framework includes a user-friendly graphical user interface (GUI) that facilitates the viewing and exporting of annotations. This feature supports real-time and pre-recorded video analysis, making SL-FE a versatile tool for sign language research.

In summary, SL-FE yields a new line of methodology of sign language phonological research. Through its theory-driven approach to phonological feature representation and analysis, SL-FE addresses the limitations of previous annotation frameworks and paves the way for new directions in sign language research and applications.

## 2. Related Works and Theoretical Aspects

### 2.1. Related Works

Traditional automatic sign language annotation frameworks have largely been oriented with a focus on feature extraction utilized in recognition and translation models for classifying handshape (Mukushev et al., 2022; Lucie Naert and Gibet, 2018), detecting sign boundaries (Momeni et al., 2022) or the recognition of lexical items (Dreuw and Ney, 2008). In the automatic annotation process, these models either utilize RGB images from sign language videos or pose estimation information in the classification of the feature set. Although these methods introduce novel architectures for automation, they heavily rely on the prior annotations done for the training. Despite the practical utility of these systems, their contribution to theoretical linguistic inquiry is less pronounced. Theoretical research on sign language linguistics, focusing on systemic structure and function, requires a detailed interpretation of sign language as a linguistic system. Recent literature reflects an increasing interest in applying pose estimation techniques to provide quantitative insights into sign languages. These studies aim to bridge the gap between signs' physical articulation and linguistic implications (Chizhikova and Kimmelman, 2022; Ghaleb et al., 2024; Keleş et al., 2023; Stamp et al., 2022). This shift has been partly propelled by advancements in pose estimation technologies, enabling the articulatory components of sign languages to be quantitatively analyzed. In response to this growing interest, our framework,

SL-FE, has been developed for both the automatic annotation of sign languages and the quantitative analysis of their phonological features concerning theoretical components of linguistic research.

### 2.2. Theoretical Aspects

Our framework's core innovation lies in its ability to provide a continuous representation of phonological features (i.e. Selected Fingers, Location, Orientation, and Movement) within a given sign language video. In the process of grounding our framework, we rely on the literature on theoretical aspects of sign language phonology where features are grouped into Inherent Features (IF) and Prosodic Features (PF) (Fenlon et al., 2017; Van der Hulst, 1993; Brentari, 1998). Namely, while Inherent Features provide a static snapshot within a single frame, the transition between position features (the thumb's interaction with the selected fingers, i.e. open to close or close to open), the transitions between settings in major locations (i.e. from proximal to distal, or from ipsilateral to contralateral), and changing orientation features (i.e. from palm to back of the hand, or from ulnar to radial parts of the hand) give rise to dynamic, Prosodic Features (PF). This treatment of phonological features and the appropriate mathematical modeling of these respective feature types are essential not only for extracting phonological information in a theoretically more informed manner from large corpora to be used in the different domains and tasks (i.e. sign segmentation and sign recognition in computer science), but they also provide a novel quantitative basis for theories of sign language phonology and typology.

## 3. Methodology

Our methodology focuses on four primary phonological feature types: Finger Selection, Orientation, Location, and Movement. Each feature type is extracted through a series of computational steps, leveraging pose-estimation technology and mathematical models to achieve a continuous and interpretable representation of sign language phonology regarding the variation and noise within sign language videos.

### 3.1. Pose Estimation

The preprocessing stage employs the Mediapipe hand and pose estimation models (Lugaresi et al., 2019), a tool for accurate human pose estimation. The model is critical to our framework, as it identifies and tracks various landmarks across the signer's body and hands in each frame, facilitating detailed phonological analysis. The landmarks include:

- **Hand Pose Landmarks:** Essential for analyzing movement, orientation, and finger selection, the model provides detailed information on the hand by identifying 21 joints per hand. Each joint is crucial for the in-depth examination of handshapes, movements, and orientations.
- **Pose Landmarks:** Primarily utilized for extracting location information, the model outputs 31 pose landmarks. These landmarks enable the framework to analyze how the signer's body interacts with space. These are either selected or generated according to the major and minor locations defined for sign languages.

Although we utilize the Mediapipe model in the current preprocessing due to its real-time processing and low CPU requirements, we are considering integrating the OpenPose framework (Cao et al., 2019). This prospective addition aims to broaden the framework's applicability and enhance its analytical depth to offer a more versatile and detailed tool for sign language research.

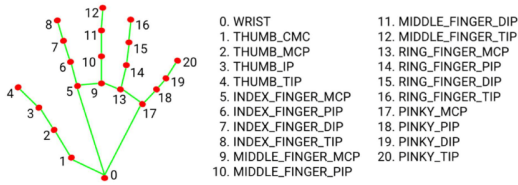


Figure 2: Hand Landmark list for left hand from Mediapipe

### 3.2. Finger Selection

Finger Selection is the first critical phonological feature our framework addresses. This process involves identifying key anchor points across each finger, focusing on four main inner joints for both hands (Joints 2-3, 6-7, 10-11, 14-15, 18-19 as designated in Figure 2). The angular distances between these joints are calculated to represent the fingers' selectional properties, such as curvature and contact points. The final feature values are obtained through min-max scaling of these angles across the video data, providing a continuous measure of finger selection within a normalized range of [0,1]. This normalization allows for a comparative analysis across different signers and sign languages, ensuring that the variations in individual signer's hand shapes do not skew the analysis.

$$FS(h, f) = \frac{1}{|J|} \sum_{p \in J} \frac{\angle(p_{j-1}, p, p_{j+1})}{180} \quad (1)$$

In the finger selection feature extraction process defined in Eq. 1,  $FS(h, f)$  serves as a quantifier

for the selection state of a given finger  $f$  on a given hand  $h$ . This mathematical representation is central to our framework, encapsulating the finger's posture in a numerical format. The set  $J$  denotes the collection of joint indices, namely, Metacarpophalangeal (MCP), Proximal Interphalangeal (PIP), and Distal Interphalangeal (DIP) joints. These joints are pivot points that define the curvature and extension of each finger.

The formula calculates the normalized average angular difference between consecutive joints in the set  $J$ . For each joint  $p$  in  $J$ , the angle  $(p_{j-1}, p, p_{j+1})$  is computed, which measures the angle at joint  $p$  formed by the line segments connecting it to its immediate neighboring joints  $p_{j-1}$  and  $p_{j+1}$ . This angle is then normalized by dividing the angle by 180 degrees to scale the value between 0 and 1. Summing these normalized angles and dividing by the cardinality of the set  $|J|$  gives us an average value,  $FS(h, f)$ , that represents the overall curvature of the finger.

The resulting feature score  $FS$  is then categorized into one of three states based on its value: "unselected" if  $FS(h, f)$  smaller than 0.2, "curved" if  $FS(h, f)$  falls between 0.2 and 0.7, indicating a partially flexed finger posture, and "selected" if  $FS(h, f)$  is greater than 0.7, signifying a finger that is actively selected by extending the finger in the formation of a sign shown in Eq. 2. This ternary categorization simplifies the interpretation of the finger's importance, distinguishing the overall handshape.

$$FS = \begin{cases} \text{unselected,} & \text{if } FS(h, f) \leq 0.2 \\ \text{curved,} & \text{if } 0.2 \leq FS(h, f) \leq 0.7 \\ \text{selected,} & \text{if } 0.7 \leq FS(h, f) \end{cases} \quad (2)$$

### 3.3. Orientation

The Orientation feature encompasses three main sub-features, each reflecting a distinct aspect of hand orientation in signing space during signing:

- **Palm-Back Score:** This score is derived from the relative orientation of the hand along the (x,y) axes, using the index knuckle and the pinky finger knuckle joints (Joints 5-17). It quantifies the extent to which the palm or back of the hand faces the interlocutor.
- **Radial-Ulnar Score:** Based on the hand's orientation along the (y,z) axes, this score also utilizes the index and pinky finger knuckle joints. It assesses the radial or ulnar deviation of the hand.
- **Tips-Wrist Score:** This score measures the orientation of the fingertips relative to the wrist along the z-axis, using the wrist and middle

finger tip joints (Joints 0-12). It captures the flexion or extension of the fingers relative to the wrist.

The granularity of phonological feature analysis is done by normalizing each orientation score to the absolute length of the signer’s hand. This axis-specific normalization ensures that the resulting scores are relative to the signer’s unique hand dimensions. Subsequently, these normalized scores are constrained within a  $[0, 1]$  range for each feature tuple. We employ a softmax function to classify these orientation labels, which provides a probabilistic interpretation of each hand orientation.

$$\hat{O}(h) = \sigma\left(\frac{\sum_{ax} |p_{ax}^1 - p_{ax}^2|}{\|p^1 - p^2\|}\right) \quad (3)$$

The equation for deriving the orientation feature vector is formulated to capture the relative position of the hand in space. In this equation,  $\hat{O}(h)$  represents the orientation feature vector for a hand  $h$ . The function  $\sigma$  denotes the softmax function, which is applied to the sum of normalized differences across a set of axes  $A$  for each feature used to define the orientation.

For each axis in  $A$ , the difference between the normalized joint positions  $p_{ax}^1$  and  $p_{ax}^2$  is calculated. These joint positions correspond to specific points on the hand, like knuckles or fingertips, relevant to the orientation being measured. The absolute value of this difference is then taken to ensure a non-negative measure of displacement. The normalization  $\|p^1 - p^2\|$  is the Euclidean distance between the two joints for each hand, serving as the denominator in the equation, which scales the orientation score relative to the size of the hand.

### 3.4. Location

Location analysis involves determining the relative positioning of each hand to major and minor locations (namely, Head, Nose, Ear, Mouth, Torso, Shoulder, and Chest). The technique measures the distance between the center of each hand and these landmarks, scaling these distances to the minimum and maximum values observed in each video frame. This scaling normalizes the data, accommodating variations in signer physique and positioning relative to the camera, thus ensuring the reliability of our phonological feature extraction across diverse datasets. We represent the overall relativized locations as the unit vector  $L$  of the distance between the center point of selection of the hand and all selected locations.

### 3.5. Movement

The Movement feature extraction is the most complex because the model synthesizes continuous

phonological information derived from each hand’s Finger Selection, Orientation, and Location analyses. Our framework models primary movement types (i.e. path movement, aperture change, and orientation change) while we are still working on modeling secondary movement types, which cannot be derived from changes in IF features (i.e. path-shape and temporal alignment properties). In this regard, although we do not provide a comprehensive movement feature set, we provide a basis for the derivation of the movement in accordance with the theoretical aspects of movement features.

To demonstrate that our framework lays a basis for deriving complex features within sign language corpora, we empirically test and display the practical implications of our model with case studies within the TID Sözlük dataset. These studies focus on phonological information complexity to substantiate theoretical claims about feature Dominance and Symmetry which we define in the next section.

## 4. Case Studies

Our framework’s application in these case studies is primarily motivated by the need to empirically test and validate phonological theories in Turkish Sign Language (TID). Utilizing the TID Sözlük dataset, we apply our framework to quantify phonological complexity to derive dominance and symmetry conditions. We have selected these two conditions regarding the theoretical discussion on these conditions indicating that the definitions are derived by difference or the similarity between information complexity between hands in two-handed signs. Earlier claims only provide hand configuration limitation on these conditions, while [Eccarius and Brentari \(2007\)](#) argue that each condition can be defined as the maximization of the difference in phonological information (Dominance) or the minimization (Symmetry) which is the initial motivation for selecting as our case studies.

### 4.1. Constraints on Two-handed Signs

The constraints on two-handed signs, concerning Dominance and Symmetry ([Battison, 1978](#)) where the Dominance Condition articulates that in two-handed signs if handshapes differ, one hand (typically the non-dominant, passive hand, or weak hand) adopts an unmarked handshape. These unmarked handshapes are typically simpler in structure. [Eccarius and Brentari \(2007\)](#) extends this by discussing featural complexity, positing a limit to the featural complexity permissible in a sign.

The study also introduces the Featural Symmetry Condition, which posits that signs reduce their featural complexity by making the two hands mirror each other regarding selected fingers and orien-

tation changes in the articulation of a sign. This suggests a balance or trade-off in complexity within the sign, resonating with the Dependency model, which views sign language structure in terms of interdependent features.

By applying these theoretical constructs to the TID dataset within our framework, we aim to provide empirical evidence for these phonological constraints. Our approach mathematically quantifies phonological complexity and symmetry, allowing us to test and validate the theoretical claims posited by phonological theory in sign language.

$$C(h) = \frac{1}{|Ch|} \sum_{ch \in Ch} abs(\Delta_h[f](ch)) \quad (4)$$

Equation 4 defines the phonological complexity  $C(h)$  for a hand  $h$  by averaging the absolute changes in phonological features across a set of channels  $C$ . In this context,  $Ch$  is a collection of channels, each representing a different aspect of phonological information, namely finger selection, orientation, and location. The function  $\Delta_h[f](ch)$  is the absolute forward finite difference function that calculates the change in a specific phonological feature  $f$  within the channel  $ch$  from adjacent frames. We obtain a measure of total phonological change by taking the absolute value of this change and summing it across all channels. This sum is then normalized by the number of channels  $|Ch|$ , resulting in an average measure of complexity for the hand across all considered phonological features. This computation allows for the quantification of complexity in a sign, providing a scalar value that can be used to analyze and compare the phonological structures within sign language corpora.

Additionally, in refining our understanding of two-handed sign constraints within the categorizations and lists the unmarked handshapes for TID [Kubuş \(2008\)](#). These definitions are particularly relevant in evaluating the performance of our framework when retrieving lexical items. The research outlines a set of unmarked handshapes specific to TID, which serve as a benchmark for assessing phonological complexity and dominance in two-handed signs.

## 4.2. Dataset

The TID Sözlük Dictionary ([Makaroğlu and Dikyuva, 2017](#)) is a comprehensive online corpus for Turkish Sign Language. It includes over 3000 isolated lexical items and within-sentence examples for each synonym. This dataset is not only a valuable educational resource but also a rich corpus for linguistic analysis, as it contains annotated handshape and location information for each lexical variant. In our study, the distribution of handshapes from this dataset serves as a basis for examining symmetry and dominance, allowing us to assign a scalar

value representing the phonological complexity for each hand.

## 4.3. Case Study on Dominance Condition in TID

Feature Dominance in sign language phonology posits that in two-handed signs, the less active hand, designated as  $h_2$ , should exhibit lower phonological complexity compared to the more active hand. This principle reflects the asymmetry often observed in the phonological structure of sign languages, where the dominant hand carries more articulatory burden.

To quantify and utilize the phonological complexity between hands in demonstrating Dominance within data, we define a ranking function for retrieving the signs that maximize the difference in complexity score.

$$\operatorname{argmax}_{H \in V} f(H) = \{\{h_1, h_2\} \in H \mid |C(h_1) - C(h_2)|\} \quad (5)$$

Equation 5 is the ranking function  $f(H)$  designed to order signs based on the maximization of phonological complexity differences between the hands. In the given sign,  $H$  represents the set containing pairs of hands, where  $h_1$  is typically the more active or dominant hand, and  $h_2$  is the less active or non-dominant hand. The function  $C(h)$  computes the phonological complexity for a given hand  $h$ .

The ranking function operates by identifying the pair of hands  $(h_1, h_2)$  within the set  $H$  that has the largest absolute difference <sup>1</sup> in phonological complexity  $|C(h_1) - C(h_2)|$ . The  $\operatorname{argmax}$  operator is applied to select the pair  $(h_1, h_2)$  for which this absolute difference is maximized across all possible hand pairs in the dictionary  $V$ . This approach inherently ranks signs in a way that emphasizes the contrast in complexity between the two hands, reflecting the dominance condition where the less active hand is expected to demonstrate less phonological complexity compared to the more active hand. The function provides a quantitative basis for ordering signs by their adherence to this phonological principle.

Investigating the Top-100 retrieved signs that exhibit the highest difference in complexity scores, we examine the distribution of handshapes for the non-dominant hand. This analysis reveals a correlation with the unmarked handshapes for TID, suggesting that less active hands tend to favor sim-

<sup>1</sup>We should note that some of the signs have the higher complexity score in left hand given dominance hands are marked general handedness of signers which is mostly right hand. It should be noted for additional studies.



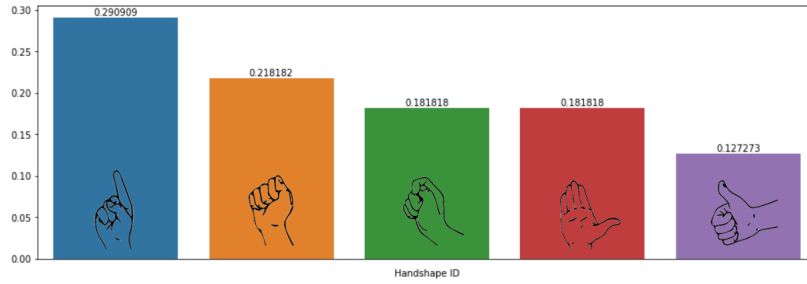


Figure 3: The handshape distribution of non-dominant hands ( $h_2$ ) for the Top-100 signs with the highest dominance ranking

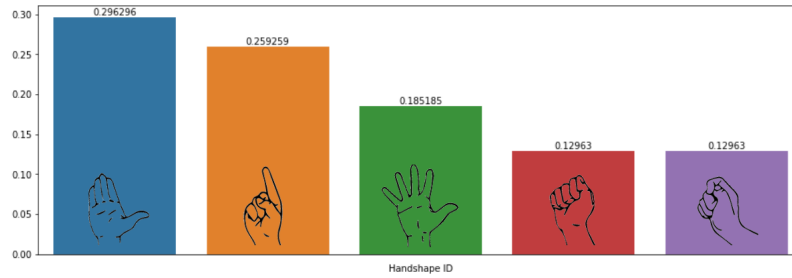


Figure 4: The handshape distribution for the Top-100 signs with the highest symmetry ranking

pler, unmarked configurations, as shown in Figure 3.

A manual annotation process assesses the accuracy of the signs retrieved by our model, resulting in a 0.90 accuracy success rate in identifying dominant-hand signs shown in Table 1. This high degree of accuracy underlines the effectiveness of our phonological complexity formulation in predicting feature dominance within the signs.

#### 4.4. Case Study on Symmetry Condition in TID

In contrast to feature dominance, the feature symmetry condition, suggests that two-handed signs should exhibit similar phonological features across both hands. This condition is motivated by featural symmetry, where both hands are expected to have similar finger selections, orientations, and movements, often resulting in unmarked handshapes.

To accommodate the symmetry condition, we revise our ranking function to focus on the minimization of phonological information complexity differences between hands as shown in Equation 6. This adjustment allows us to evaluate the degree of symmetry in the phonological structure of each sign by identifying and prioritizing those with the least complexity difference between the hands.

$$\operatorname{argmin}_{H \in L} f(H) = \{ \{h_1, h_2\} \in H \mid |C(h_1) - C(h_2)| \}$$

(6)

Following the re-ranking of signs according to the updated function, we investigate the distribution of handshapes, particularly looking for the occurrence of unmarked shapes that would be indicative of symmetry. We then assess the accuracy of our model's ability to detect symmetric signs. A higher accuracy rate in this assessment would support our framework's capability to model phonological complexity effectively and validate the feature symmetry condition in sign language phonology. Similar to the Dominance Condition, we also observed the high distribution of unmarked handshapes in Top-100 retrieved sign as shown in Figure 4.

#### 4.5. Results

In the dominance condition analysis, the model demonstrated high performance. This accuracy is attributed to the framework's capability to maximize the phonological information differences between the hands, which is a direct quantification of the dominance condition. The results were consistent with theoretical expectations, affirming the model's validity in discerning the more active hand's increased complexity. While still accurate, the analysis of the symmetry condition revealed lower performance metrics compared to the dominance model. This outcome is due to the complexity of symmetry, which is not solely about minimizing differences between hands but each hand should yield lower complexity separately. This dual requirement supports theoretical assertions of Eccarius

and Brentari (2007) and highlights the additional constraints involved in modeling symmetry within sign language phonology. Nevertheless, the baseline scores provided for the retrieval of symmetry are still relatively high for further studies.

Label	Acc.	Pre.	Rec.	F1
Dominance	0.90	1.00	0.90	0.95
Symmetric	0.84	0.71	0.84	0.77

Table 1: The results of the performance of retrieved Top-100 signs with highest Dominance and Symmetric ranking

## 5. Future Work

Further developments in our framework will address the integration of movement features, which are dynamic and complex components of sign language. We plan to utilize neural network models to effectively model these features, which can learn and generalize from large datasets. These models can potentially capture the temporal and spatial movement information across sign languages, translating them into meaningful phonological data that can be used for further linguistic analysis.

The ultimate goal of our research is to achieve a fully automated annotation process for sign language videos via advanced neural models. This automation will not only accelerate the annotation process but also enhance its accuracy, consistency, and scalability. As we integrate these advanced neural models, we will also re-evaluate and refine our annotation methodologies to ensure they remain robust and reliable for comprehensive sign language research.

## 6. Conclusion

In conclusion, SL-FE proves to be a transformative tool for sign language phonological analysis, adeptly bridging the gap between theoretical models and practical annotation. It offers a novel computational approach to quantify phonological complexity, providing empirical evidence for longstanding theoretical constructs. The case studies conducted with the TID dataset affirm the framework's capability to identify feature dominance and symmetry. Moreover, applying the two-handed sign criteria confirms the phonological constraints and others posited. As we continue to refine SL-FE, we anticipate its broader application in sign language research.

## 7. Acknowledgment

We would like to thank Lale Akarun and Umit Atlamaz for their contribution to reviewing the mathematical aspects of our framework and providing the necessary feedback and improvements.






## 8. Bibliographical References

- Kairat Aitpayev, Shynggys Islam, and Alfarabi Ima-shev. 2016. [Semi-automatic annotation tool for sign languages](#). In *2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4.
- Robbin Battison. 1978. Lexical borrowing in american sign language.
- Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Brafort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. [Sign language recognition, generation, and translation: An interdisciplinary perspective](#). *CoRR*, abs/1908.08597.
- Diane Brentari. 1998. *A prosodic model of sign language phonology*. Mit Press.
- Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Anastasia Chizhikova and Vadim Kimmelman. 2022. [Phonetics of negative headshake in Russian Sign Language: A small-scale corpus study](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 29–36, Marseille, France. European Language Resources Association.
- Philippe Dreuw and Hermann Ney. 2008. [Towards automatic sign language annotation for the ELAN tool](#). In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 50–53, Marrakech, Morocco. European Language Resources Association (ELRA).
- Petra Eccarius and Diane Brentari. 2007. [Symmetry and dominance: A cross-linguistic study of signs and classifier constructions](#). *Lingua*,

- 117(7):1169–1201. The linguistics of sign language classifiers: phonology, morpho-syntax, semantics and discourse.
- Jordan Fenlon, Kearsy Cormier, and Diane Brentari. 2017. The phonology of sign languages. In *The Routledge handbook of phonological theory*, pages 453–475. Routledge.
- Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Asli Özyürek, and Raquel Fernández. 2024. Co-speech gesture detection through multi-phase sequence labeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4007–4015.
- Matilde Gonzalez, Michael Filhol, and Christophe Collet. 2012. [Semi-automatic sign language corpora annotation using lexical representations of signs](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2430–2434, Istanbul, Turkey. European Language Resources Association (ELRA).
- Marek Hruź, Zdeněk Krňoul, Pavel Campr, and Luděk Müller. 2011. Towards automatic annotation of sign language dictionary corpora. In *International Conference on Text, Speech and Dialogue*, pages 331–339. Springer.
- Onur Keleş, Ceren Oksal, and Emre Bilgili. 2023. [Using pose estimation for reference tracking in turkish sign language](#). In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4.
- Okan Kubuş. 2008. An analysis of turkish sign language (tid) phonology and morphology. Master's thesis, Middle East Technical University.
- Caroline Larboulette Lucie Naert, Clément Reverdy and Sylvie Gibet. 2018. Per channel automatic annotation of sign language motion capture data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. [Mediapipe: A framework for building perception pipelines](#).
- Bahtiyar Makaroğlu and Hasan Dikyüva. 2017. *Güncel Türk İşaret Dili Sözlüğü*.
- Liliane Momeni, Hannah Bull, KR Prajwal, Samuel Albanie, Gül Varol, and Andrew Zisserman. 2022. Automatic dense annotation of large-vocabulary sign language videos. In *European Conference on Computer Vision*, pages 671–690. Springer.
- Medet Mukushev, Arman Sabyrov, Madina Sultanova, Vadim Kimmelman, and Anara Sandygulova. 2022. [Towards semi-automatic sign language annotation tool: SLAN-tool](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 159–164, Marseille, France. European Language Resources Association.
- Sylvie C. W. Ong and Surendra Ranganath. 2005. [Automatic sign language analysis: A survey and the future beyond lexical meaning](#). *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):873–891.
- Victor Skobov and Yves Lepage. 2020. [Video-to-HamNoSys automated annotation system](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 209–216, Marseille, France. European Language Resources Association (ELRA).
- Rose Stamp, Lilyana Khatib, and Hagit Hel-Or. 2022. [Capturing distalization](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 187–191, Marseille, France. European Language Resources Association (ELRA).
- Harry Van der Hulst. 1993. Units in the analysis of signs. *Phonology*, 10(2):209–241.

# Signs and Synonymity

## Continuing Development of the Multilingual Sign Language Wordnet

Marc Schulder<sup>1</sup>, Sam Bigeard<sup>1,2</sup>, Maria Kopf<sup>1</sup>, Thomas Hanke<sup>1</sup>,  
Anna Kuder<sup>3</sup>, Joanna Wójcicka<sup>4</sup>, Johanna Mesch<sup>5</sup>, Thomas Björkstrand<sup>5</sup>,  
Anna Vacalopoulou<sup>6</sup>, Kyriaki Vasilaki<sup>6</sup>, Theodore Goulas<sup>6</sup>,  
Stavroula–Evita Fotinea<sup>6</sup>, Eleni Efthimiou<sup>6</sup>

<sup>1</sup>Institute of German Sign Language and Communication of the Deaf,  
University of Hamburg, Germany

<sup>2</sup>Inria Centre, University of Lorraine, France

<sup>3</sup>Department of Linguistics, University of Cologne, Germany

<sup>4</sup>Department of General Linguistics, Sign Language Linguistics and Baltic Studies,  
University of Warsaw, Poland

<sup>5</sup>Department of Linguistics, Stockholm University, Sweden

<sup>6</sup>Institute for Language and Speech Processing, Athena Research Center, Greece  
{marc.schulder, maria.kopf, thomas.hanke}@uni-hamburg.de, sam.bigeard@inria.fr,  
akuder@uni-koeln.de, j.filipczak@uw.edu.pl, {johanna.mesch, bjorkstrand}@ling.su.se,  
{avacalop, kvasilaki, tgoulas, ndimou, evita, eleni\_e}@athenarc.gr

### Abstract

The Multilingual Sign Language Wordnet is the first publicly available wordnet resource for sign languages. It is a growing multilingual resource providing data for eight sign languages to date. During the initial phase of its creation, the focus lay on producing the infrastructure to support various languages and to produce initial sets of content for them. This article represents the start of the second phase, in which the focus is moved to establishing overlapping coverage across the different sign languages. Building on the data produced so far, a new feature to assist annotation is introduced which leverages established partial synonymy between signs (inter- and cross-lingually) to discover likely additional synonymies. Other improvements to the annotation interface and workflow build directly on the experiences from the first phase. Working with the updated annotation interface, new data is produced for Polish Sign Language, Greek Sign Language and Swedish Sign Language.

**Keywords:** multilingual wordnet, sign language wordnet, resource creation, dictionary reversal

## 1. Introduction

The *Multilingual Sign Language Wordnet (MSL-WN)*<sup>1</sup> is the first publicly available wordnet resource for sign languages. It connects the inventory of several lexical sign language resources with the synset inventory of Open Multilingual Wordnet (OMW) (Bond et al., 2016). It so far provides data for eight different sign languages:

- British Sign Language (BSL)
- German Sign Language (DGS)
- Swiss German Sign Language (DSGS)
- Greek Sign Language (GSL)
- French Sign Language (LSF)
- Sign Language of the Netherlands (NGT)
- Polish Sign Language (PJM)
- Swedish Sign Language (STS)

The data of *MSL-WN* is of use both for sign language processing, e.g. to counter data sparsity issues in machine translation, and for linguistic research, especially in cross-lingual studies. For resource creators it provides an opportunity to connect different resources on a semantic level, overcoming compatibility issues caused by differences in e.g. glossing practices (Kopf et al., 2022b).

Until the end of 2023, *MSL-WN* was created in the context of the EU project EASIER<sup>2</sup>. The focus of this first phase was on establishing basic coverage for a large number of different sign languages. Lexical material for sign languages and language expertise for annotation were established through collaborations with numerous data owners. Workflows were directly informed by what data was available and what was not, leading to a number of trade-offs. Automatic assistive methods relied solely on spoken language data, requiring annotators to counter-act errors introduced by translation, differences between spoken and signed modality, and the automatic matching algorithm. Annotation for different languages also mostly worked in isola-

<sup>1</sup><https://doi.org/10.25592/dgs.wn>

<sup>2</sup><https://doi.org/10.3030/101016982>



tion from each other, as cross-lingual information was scarce, although annotators were encouraged to prioritise the verification of candidates for synsets that already had links for other languages. Nevertheless, only 16% of synsets were linked to more than one sign language.

This paper represents the start of the second phase for MSL-WN, introducing a special focus on improving cross-lingual coverage between sign languages. It is time to revisit and optimise the established workflows and to consider how the data that has been produced so far can be used to further support the annotation process. The rest of this article presents relevant related work (Section 2) and a summary of the MSL-WN creation process so far (Section 3), followed by a discussion of the lessons learned (Section 4). Section 5 introduces a new automatic suggestion feature that relies on direct (partial) synonymy between signs instead of relying on translation to a spoken language. The article is also accompanied by a new release of the MSL-WN dataset, providing new annotations for PJM, STS and GSL that were produced with the assistance of the new enhanced annotation interface (Section 6). We conclude in Section 7 with an outlook on future steps.

## 2. Related Work

This section discusses relevant related work regarding spoken language wordnets (Section 2.1), sign language wordnets (Section 2.2) and approaches to support the creation of lexical resources (Section 2.3).

### 2.1. Wordnets for Spoken Languages

The first wordnet was the Princeton Wordnet (PWN) for English (Fellbaum, 1998) and it is still the most complete and widely used wordnet in existence. It was introduced by Miller et al. (1990) as a psycholinguistically motivated alternative to the traditional approach of organising dictionaries by the alphabetical order of citation forms. Words are grouped into synsets, sets of synonyms, each of which represents a specific concept. The different senses of a polysemous word are expressed through its inclusion in several synsets. The result is a many-to-many network of forms and meanings. Furthermore, synsets are connected to each other through hyponymy and other relations, creating a taxonomic hierarchy that represents the main organisational structure of the wordnet.

Following the example of PWN, wordnets for various languages (mainly spoken languages with conventionalised written forms) have been developed since (Bond and Paik, 2012). Several projects, such as EuroWordNet (Vossen, 1998), BalkaNet

(Tufiş et al., 2004) and African WordNet (Le Roux et al., 2008), have also worked on creating aligned wordnets for several languages. Many wordnets with open access licences have since been connected into an interconnected network of wordnets by the Open Multilingual Wordnet (OMW) (Bond and Paik, 2012).

### 2.2. Wordnets for Sign Languages

A number of reports on creating sign language wordnets exist. Work on individual sign languages was reported for DSGS (Ebling et al., 2012), Italian Sign Language (LIS) (Shoaib et al., 2014) and American Sign Language (ASL) (Lualdi et al., 2021), although the data was not made publicly available at the time. In all cases, the authors made use of existing lexical resources, allowing them to leverage available lexical information and video material to drastically reduce production cost and in turn provide added value to those lexical resources. Lualdi et al. (2021) also reported on combining several resources to increase the available vocabulary.

Other works use wordnets to support software functions or internal work processes. The DictaSign project (Matthes et al., 2012) defined a list of 1,000 concepts, each represented by a PWN synset, for which they provided signs in four languages: BSL, DGS, GSL and LSF (Dicta-Sign Consortium, 2012). The synsets were used to connect signs cross-lingually, provide concept definitions and allow synonym-based spoken language text search through the project's web interface (Efthimiou et al., 2012).

The *Danish Sign Language Corpus and Dictionary* project (Troelsgård and Kristoffersen, 2018b) link their sign type inventory to synsets from DanNet (Pedersen et al., 2009) to enhance the annotator's type search by also matching to Danish synonyms (Troelsgård and Kristoffersen, 2018a). Declerck and Olsen (2023) reported on-going work on making this information publicly available as linked open data.

Langer and Schulder (2020) automatically match lexical entries of the DGS Corpus (Prillwitz et al., 2008) with lemmas from GermaNet (Hamp and Feldweg, 1997) to extract supersense categories for use in coarse semantic clustering for lexicographic work, although no sense disambiguation is performed.

The DSGS data by Ebling et al. (2012) and the LSF data of the DictaSign project have been integrated into the MSL-WN (Bigéard et al., 2024).

### 2.3. Bootstrapping lexical content

Creating lexical resources is labour intensive and many methods to support or automate this work have been considered. For bilingual dictionaries, a

common technique (these days aided by the use of lexicographic editing software) is that of *dictionary reversal*, in which one translation direction is produced first and then a first draft of the inverse direction is created by reversing the entries (cf. [Martin, 2013](#)). As [Martin \(2013\)](#) points out, relying purely on the surface forms of words for reversal can produce many mistranslations caused by incorrectly mapping the different meanings of polysemous words across languages. More robust reversals can be achieved when building on lexical units that represent individual meanings (naturally, additional complications still arise, see for example [Corda et al., 1998](#)). [Lam and Kalita \(2013\)](#) leverage the concept inventories of wordnets to this end, providing an algorithm for dictionary reversal that relies on having wordnet data for one of the two languages involved.

For the creation of new wordnets, a major hurdle is producing the required synset inventory. Many projects follow the *expand model* ([Vossen, 1998](#), p. 83) in which an existing wordnet, usually PWN, is used as a foundation upon which to expand, significantly reducing the required amount of work ([Bosch and Griesel, 2017](#)). As a side effect, wordnets that expand from the same wordnet also acquire cross-lingual compatibility through their shared concept inventory. Building on this idea, [Bond et al. \(2016\)](#) introduced the Collaborative InterLingual Index (CIL), an extension of the PWN synset inventory that allows consistent identification of synsets and addition of new synsets and relations to account for concepts and linguistic structures missing from English or Anglocentric cultures. CIL is directly integrated into OMW ([Vossen et al., 2016](#)).

### 3. Creating the Multilingual Sign Language Wordnet

The MSL-WN was started by the EASIER project as a publicly available cross-lingual semantic resource for use in sign language technologies. This section describes work that happened up until the end of the project in December 2023, also documented in project report D6.5 ([Bigéard et al., 2024](#)).

During the project, it received three releases: An initial proof-of-concept release providing data for GSL and DGS ([Bigéard et al., 2022](#)), a second release covering the remaining project languages (BSL, DSGS, NGT, and LSF) ([Bigéard et al., 2023](#)) and a final release introducing the project-external languages PJM and STS ([Bigéard et al., 2024](#)).

Inclusion of STS and PJM was made possible through partnerships with the creators of the Swedish Sign Language Dictionary ([Svenskt teckenspråkslexikon, 2024](#)) and the Corpus-based Dictionary of Polish Sign Language ([Łacheta et al., 2016](#)), who are also co-authoring this article. Simi-

larly, video material and lexical information for the other languages were taken from existing lexical resources, each of which is credited by MSL-WN, including reference links for each individual sign entry in the MSL-WN web interface.

#### 3.1. The annotation interface

To support the annotation efforts for MSL-WN, a custom web interface was developed. It is regularly updated to accommodate new data, add additional features and react to annotator feedback (see [Section 4](#)).

The interface provides two annotation perspectives: in *sign view* all meanings (i.e. synsets) of a single sign are annotated, while in *synset view*, one synset is associated with all signs that can represent its meaning. Indices for either perspective provide additional filters, such as listing only synsets that have already been annotated for other sign languages as well, to help annotators focus on what to annotate next. Where frequency information is available from the underlying lexical resource, it is used to sort content, so that more commonly used signs are annotated first. A demonstration of the *sign view* interface is shown in [Figure 1](#). For further information regarding the annotation workflow and interface, see [Bigéard et al. \(2022\)](#) and [Bigéard et al. \(2024\)](#).

#### 3.2. Gloss-based suggestions

To reduce the required amount of manual search for sign-synset connections, the annotation interface provides automatically determined suggestions of likely candidate connections. These suggestions can be generated using different methods. This section discusses the first method, *gloss-based suggestions*<sup>3</sup>, introduced in [Bigéard et al. \(2022\)](#), while a new second method, *synonym-based suggestions* is introduced in [Section 5](#).

Due to the lack of established written forms for sign languages, sign language resources commonly supplement the video representation of signs with ID-glosses or keyword translations to textually represent them in lexicon entries, annotations and search interfaces. These are most often produced in the dominant spoken language of the geographic region of the sign language, although some projects also produce additional English versions to facilitate international exchange.

Gloss-based suggestions leverage this information by matching glosses and keywords to lemmas


---

<sup>3</sup>For brevity, we use the term *gloss-based suggestion* regardless of exactly which spoken language representation is provided for a sign. Depending on the underlying lexical resource, the used text may be a one or several glosses, translational equivalents or other forms of spoken language keyword.

## Manual Annotation for the Multilingual Sign Language Wordnet

Home | Browse STS | Browse PIM | Browse GSL | Browse DGS | BSL | DSGS | LSP | NGT | Interlingual synsets | EASIER synsets | All synsets

pjm.1257 leżec/polożyć 2 [external link](#)



pracownia lingwistyki migowej

0:02 / 0:04

You are annotating whether the synsets below should be linked to the sign above.  
[Search for a missing synset](#)

Click to mark this sign as not correct for ALL synsets below that have NOT YET been validated. This is the same as individually clicking "not correct" on all the synsets displayed below that you have not worked on yet, but quicker.

Mark not correct everywhere else

Synset ID and links	Synset info	Other signs with that meaning	Validation Status
Automatic suggestions based on <a href="#">Gloss-to-Word Auto-Match (weak)</a>			
<a href="#">omw.00834259-v</a> <a href="#">view in OMW</a>	<b>English lemmas:</b> lie <b>Definition:</b> tell an untruth; pretend with intent to deceive <b>Examples:</b> <ul style="list-style-type: none"><li>• Don't lie to your parents</li><li>• She lied when she told me she was only 29</li></ul>		validate at <input type="checkbox"/> ok! <input checked="" type="checkbox"/> correct: this sign can mean this <input type="checkbox"/> not correct: this sign cannot mean this ... I'm not sure, mark for review
validated as correct			
<a href="#">omw.01494310-v</a> <a href="#">view in OMW</a>	<b>English lemmas:</b> put, set, place, pose, position, lay <b>Definition:</b> put into a certain place or abstract location <b>Examples:</b> <ul style="list-style-type: none"><li>• Put your things here</li><li>• Set the tray down</li><li>• Set the dogs on the scent of the missing children</li><li>• Place emphasis on a certain point</li></ul>	<a href="#">pjm.3226</a> <a href="#">pjm.1589</a> <a href="#">gsl.539</a> <a href="#">bsl-dictasign.697</a> <a href="#">lsf.697</a> <a href="#">gsl-dictasign.697</a> <a href="#">dgs-dictasign.697</a>	validated as correct <input checked="" type="checkbox"/> <a href="#">undo validation</a>
<a href="#">omw.02690708-v</a> <a href="#">view in OMW</a>	<b>English lemmas:</b> lie <b>Definition:</b> be located or situated somewhere; occupy a certain position	<a href="#">pjm.1521</a> <a href="#">pjm.1258</a>	validated as correct <input checked="" type="checkbox"/> <a href="#">undo validation</a>
validated as wrong			
<a href="#">omw.06756831-n</a> <a href="#">view in OMW</a>	<b>English lemmas:</b> lie, prevarication <b>Definition:</b> a statement that deviates from or perverts the truth	<a href="#">pjm.2301</a> <a href="#">pjm.2328</a> <a href="#">bsl-dictasign.491</a> <a href="#">lsf.491</a> <a href="#">dgs-dictasign.491</a> <a href="#">gsl-dictasign.491</a>	validated as wrong <input checked="" type="checkbox"/> <a href="#">undo validation</a>
<a href="#">omw.11131808-n</a> <a href="#">view in OMW</a>	<b>English lemmas:</b> Lie, Trygve_Lie, Trygve_Halvdn_Lie <b>Definition:</b> Norwegian diplomat who was the first Secretary General of the United Nations (1896-1968)		validated as wrong <input checked="" type="checkbox"/> <a href="#">undo validation</a>

Figure 1: *Sign view* perspective of the MSL-WN annotation interface. The video, ID, glosses and keywords are shown at the top, followed by features for searching missing synsets and mass-rejecting automatic suggestions. The table below lists synsets that are either candidates suggested by automatic systems or have been validated by a human annotator, grouping them accordingly. For each synset, annotators are given links to its entries in the annotation interface and in OMW, lemmas, definitions and examples from available spoken language wordnets, and a list of other signs from the same and other languages already linked to the synset. *Note that the above screenshot has been abridged to allow inclusion in this article.*

found in a wordnet for that spoken language and returning their synsets as candidates. There are several limitations to this approach. As a variant of form-based reverse translation (see [Section 2.3](#)), it tends to over-generate for polysemous words, providing many senses that do not apply to the target sign. At the same time, in cases where the lexical data of the sign only provides a single form-level gloss (commonly a best-effort translation for what is assumed the most dominant sense of the sign), secondary senses are still missed out on. Additional technical challenges come from converting glosses to lemmas ([Kopf et al., 2022b](#)), handling complex expressions and processing abbreviations and lexicographic addenda.

Nevertheless, given the limited available data and lack of language technologies to support alternative approaches, gloss-based suggestions still represented the best assistive method viable at the time. It was also clearly preferable to requiring annotators to manually look up each synset, a con-

siderably slower approach with its own pitfalls and which in the end also relies on spoken language lemma lookup. Further observations on how gloss-based suggestions affected the annotation process are provided in [Section 4](#).

## 4. Experiences with Annotation

In this section we share some of the experiences of annotation teams during phase 1 of MSL-WN and how these affected the continuing development of the annotation interface.

### 4.1. Accommodating different workflows

As was mentioned in [Section 3.1](#), the annotation interface allows annotators to focus on annotating either a specific sign or a specific synset through different annotation views and index lists. Both views were adopted by annotators, with individual

annotators showing different preferences for focusing on the *sign view*, *synset view* or dynamically switching between both.

Focusing on a specific view will, of course, have an impact on how the coverage of the dataset progresses, with work through *sign view* contributing to more complete description of all senses of a sign, and *synset view* aiding the interconnectedness of signs for a given concept, although both approaches should largely converge in coverage as annotation progresses. Nevertheless, annotators are encouraged to switch between views, both during exploration of the data and between completed annotation passes.

## 4.2. Language competence

Language competence of annotators played an important role both regarding the language being annotated and the language that wordnet data is available in.

### 4.2.1. Annotation language

Ideally, annotators should have L1 competence in the sign language they are annotating. However, given the high demand for L1 signers in linguistic research, some language teams had to fall back on fluent L2 signers as annotators. In these cases, the L2 signers took care of the majority of annotations, consulting the lexical and corpus data of the resource from which MSL-WN took the sign entry, but deferred unclear cases to an L1 signer. The annotation interface was designed to accommodate this need by letting annotators mark entries as needing review, rather than being valid or invalid.

### 4.2.2. Synset description language

To determine the meaning of a synset, annotators can reference its definition, its list of example sentences and the set of words and signs associated with it. This means that annotators need to be competent in the languages in which this information is provided. While in some cases synset definitions or examples in multiple (spoken) languages are available, often only information in English is available. This can pose challenges for annotators, particularly regarding specific nuances of meaning between closely related synsets, and can add additional requirements regarding the multilingual competence of annotators.

Disambiguating information may also be less complete for some languages. For instance, the GSL team found that Greek synset definitions from BalkaNet were often missing usage examples. This resulted in a degree of uncertainty on the part of annotators as to the accuracy of their choices for sign-synset connections.

At the same time, annotators reported that seeing what signs from other languages were already assigned to a synset could be very helpful when the annotator had competence in that other language as well. This also (slightly) helps to mitigate the issue that definitions are only ever available in spoken languages.

### 4.2.3. Languages for gloss-based suggestions

Availability of languages also played a big role in how well gloss-based suggestions could be. Most lexical resources provide their text information in the dominant spoken language of the sign language's region. English descriptions may also be provided by some resources, but these are often secondary translations intended to widen access for the international research community.

Where possible, gloss-based suggestions use the regional spoken language. However, as the size of different wordnets can differ strongly and none rival that of PWN, English often has to be used as a fallback option to generate any suggestions at all. For example, only 30% of suggestions for PJM could be generated via Polish keywords and wordnet entries, while 70% were based on English.

Furthermore, verifying the quality of the automated matches between glosses/keywords and wordnet lemmas often fell on annotators as well, as the developers of the matching procedures did not necessarily have the required language competence to do so.

## 4.3. Changing workflows for GSL

The very first data produced for MSL-WN were annotations of GSL and DGS. During this initial experimental period, different possible workflows were explored in parallel (see [Bigéard et al., 2022](#)). As the MSL-WN annotation interface was not yet available, the GSL team used its own internal lexicographic workflows and software (for details, see [Vacalopoulou et al., 2022](#)).

Annotation focused on the Greek language part of OMW, a set of 18,000 synsets originally produced for BalkaNet ([Tufiş et al., 2004](#)). Using this data allowed the team to produce gloss-based automatic suggestions between Greek wordnet lemmas and the Greek keywords of the GSL lexical database Noema+ ([Efthimiou et al., 2016](#)). It also allowed annotators to work with the written language for which they had the strongest language competence, although some issues regarding completeness of wordnet information occurred (see [Section 4.2.2](#)). In addition, the known flaws of gloss-based suggestions discussed in [Section 3.2](#) resulted in invalid sign-synset link suggestions in about 30% of the cases.



For the second phase of MSL-WN, the GSL team switched to working with the MSL-WN annotation interface. Adjustment to the new interface was found to be straightforward with no major issues. Having all relevant information gathered in one interface, rather than cross-referencing independent resources, reportedly helped annotators to stay focused on their task. Being shown sense-level synonyms from other (sign) languages (see [Figure 1](#)) also helped annotators with knowledge of those languages to distinguish between possible senses (see [Section 4.2.2](#)).

#### 4.4. Evolving the interface

The MSL-WN annotation interface and underlying data structures evolved continuously as annotation progressed. For each new language, custom procedures were developed to transfer information from the underlying lexical resource to MSL-WN. The computation of gloss-based suggestions also had to be adjusted to both the language being processed and the structure of lexical resource data (see [Section 3.2](#)).

Development of the interface directly took into account annotator feedback. For example, initially, each sign-synset link, including each automatic suggestion, had to be validated individually. Based on a request by the NGT annotation team, a button to summarily reject all automatic suggestions of a specific sign or synset was added ([Bigéard et al., 2024](#)). This allows annotators to first validate correct suggestions and then reject the remaining suggestions in one go, speeding up the annotation process. This functionality is currently being extended further to generally allow the dynamic selection of multiple entries for joint submission of a shared validation decision.

#### 4.5. Expanding the sign inventory

The sign inventory of MSL-WN is defined by the lexical sign language resource on which it builds. Accordingly, it inherits any limitations that the lexical resource might have. For instance, the vocabulary of the *Corpus-based Dictionary of Polish Sign Language* ([Łacheta et al., 2016](#)) is determined by what signs could be observed in the *Polish Sign Language Corpus* ([Kuder et al., 2022](#)). The content of the corpus was in turn affected by its selection of elicitation tasks, so that certain topics are more prevalent than others. As a result, neither corpus nor dictionary cover e.g. slang terms or specialised terminology.

Limitations of vocabulary can directly affect the work of annotators, who might be aware of signs existing for a synset sense, but need to verify through lengthy searches whether it is part of the resource or not.

The best way to address missing vocabulary is, of course, to introduce it. This would either require the production of new lexical materials (which is better achieved by our lexical resource partners than MSL-WN itself) or the inclusion of additional lexical resources. Apart from the general rarity of such resources (see [Kopf et al., 2022a](#)), this also introduces the question of how best to identify overlaps in sign inventory between resources of the same language, so as to avoid creating duplicate sign entries.

### 5. Leveraging Sign-to-Sign Synonymy

Since its inception, the MSL-WN has grown to include over 10,000 verified sign-synset links. To leverage this data for future annotation work, we introduce a new method for generating automatic suggestions of possible sign-synset links. This new method suggests additional meanings based on the sense inventories of other signs that have already been verified to be (partial) synonyms of the sign being annotated. This method can be applied both intra- and cross-lingually. It represents a reversal based on lexical units, which is preferable to the form unit reversal of gloss-based suggestions (see [Sections 2.3](#) and [3.2](#)). It also reduces the dependency on spoken languages as a pivot.

#### 5.1. Implementing synonym suggestions

Synonym-based suggestions are determined as follows: Once two signs are established as having partial synonymy, i.e. they have at least one shared meaning, expressed through both signs being linked to the same synset, there is a reasonable chance that they also share other meanings. When one of the signs is being annotated, the other sign is checked for verified links to additional synsets, which can then be provided as suggestions for the current annotation. A concrete example of this procedure is shown in [Figure 2](#).

In the annotation interface, synonym-based suggestions are grouped separately from gloss-based suggestions and ranked higher (cf. groupings in [Figure 1](#)). For sorting the different synonym-based suggestions amongst themselves, the following ranking steps are applied:

1. **Intra- or cross-lingual:** Suggestions are possible both between signs of the same language and across languages, but those from the same language are ranked more highly. Connections that are only established cross-lingually are grouped separately.
2. **Synonym purity:** Apart from annotating valid sign-synset links, annotators can also explicitly mark invalid relations, meaning that a sign

## Current annotation task

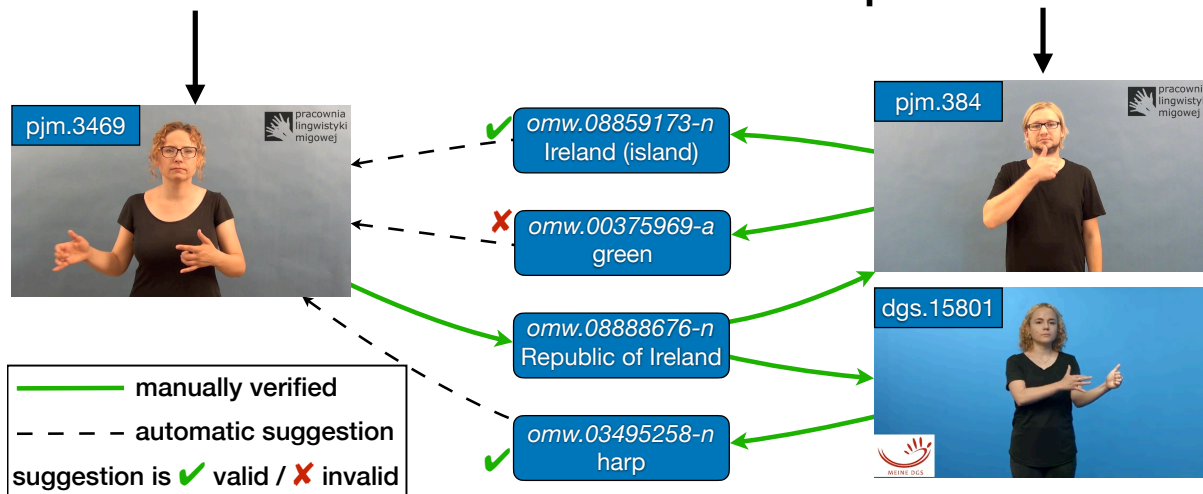


Figure 2: Demonstration of automatic suggestions based on partial synonymy between signs. The sign `pjm3469` is currently being annotated. Once at least one shared meaning is established between signs (here via synset `omw.08888676-n`), the interface suggests other possible shared meanings based on the verified annotations of the other signs. In this case it correctly suggests synsets representing the island of Ireland and a harp, but incorrectly suggests the concept of the colour green. (Image sources: Łacheta et al., 2016; Konrad et al., 2020)

does *not* have the meaning represented by this synset. As a result there can be sign pairs that are confirmed to be only partially synonymous, because while both have valid links to some synsets, there are other synsets that are marked as valid senses of one sign and invalid senses of the other sign. These partial synonyms are ranked lower than signs whose synonymy is (so far assumed to be) more complete.

3. **Synonymy strength:** Sign pairs with established synonymy across several senses are more likely to be fully synonymous, so suggestions are ranked higher the more verified synset links a sign pair has in common.
4. **Sign-to-sign quantity:** If a synset is suggested several times through synonym connections with different signs, this candidate is ranked higher than if only one connection suggested it.

These ranking steps are not without flaws, as they can be impacted by how complete the annotations of individual other signs are as well as structural factors of the used sign inventories. For example, phonological variants of a sign are often listed as separate entries, which could inflate their weight in the *sign-to-sign quantity* ranking. Additional information from the underlying lexical resources may be used in future to partially counteract such issues.

## 5.2. Preliminary annotator feedback

Annotators recently started working with the new synonym-based suggestion feature. Initial observations were that it clearly helps discover senses that were not covered through the spoken language pivot of gloss-based suggestions, in some cases significantly increasing sense coverage of polysemous signs.

It also speeds up annotation of form and dialectal variants, which are usually represented as separate sign entries, but are largely (though not necessarily fully) synonymous. Annotators also welcomed having a more sign-centric approach that helps reduce reliance purely on spoken language data.

In some cases, signs receive a large number of incorrect suggestions when they are linked to a highly polysemous sign. To quickly handle such cases, we are investigating interface improvements to allow a focused inspection of suggestions based on what connection triggered them.

MSL-WN has always kept a record not only of valid sign-synset connections, but also of which connections were verified as incorrect. This is becoming particularly useful in connection with synonym-based suggestions, as it provides hard data on the scope and limits of partial synonymy between signs. This is expected to help linguistic research, for instance in studies comparing how strong the actual synonymy between signs from different languages is.

Published	GSL	PJM	STS	Total
Signs	2,949	2,415	2,706	14,028
Synsets	5,760	3,626	1,981	16,534
Links	7,499	6,486	2,810	24,367

Table 1: Counts of signs, synsets and sign-synset links included in the public MSL-WN dataset. Shows the languages currently being worked on and the sums for all languages.

## 6. New Dataset Release

While the EASIER project has ended, work on MSL-WN continues. The first step is the annotation of additional data for PJM, STS and GSL. The current state of this on-going effort has already been published to the MSL-WN dataset.<sup>4</sup> Statistics for recently updated languages can be seen in Table 1.

This new round of annotations is produced in concert with the interface and workflow improvements presented in this article. Supported by the new sign-to-sign synonym suggestion feature, annotators have been encouraged to focus on synsets that have already been verified for other signs. Compared to the release described in Bigeard et al. (2024), the number of synsets covered by at least two languages rose by 79% to 3,361 and those covered by at least three or four languages rose by 59% (1,472) and 51% (617), respectively.

## 7. Conclusion

This article presents the current progress of the Multilingual Sign Language Wordnet (MSL-WN). It describes the experiences gathered so far and how workflows and software support were adjusted according to annotator feedback. A major change is the recent addition of a new automatic suggestion feature that leverages established partial synonymy between signs. This helps reduce the dependence on spoken language form-level suggestions.

Annotation of Multilingual Sign Language Wordnet (MSL-WN) is on-going. A new dataset release which introduces new annotations for GSL, PJM, and STS accompanies this article. Work on these annotations is also used to gather initial feedback on the impact of the new interface features.

In future, a varied number of improvements are planned. Some of these regard the fluency of the annotation workflow, such as allowing more flexible simultaneous verification of sets of suggestions. Others will address questions of how to integrate multiple lexical resources for one sign language and how to represent sign language phenomena such as classifiers and incorporation in wordnet.

<sup>4</sup><https://doi.org/10.25592/uhhfdm.14190>

## 8. Acknowledgements

We would like to thank Neil Fox, Kearsy Cormier, Onno Crasborn, Lianne Westenberg, Sarah Ebling and Laure Wawrinka for their work in prior phases of MSL-WN on which this article builds. We would like to thank Gabriele Langer for valuable discussions regarding lexicographic practices.

This publication has been produced in part in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the Academies of Sciences and Humanities.

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982.

This work is supported in part by INRIA within the project Defi COLaF.

AK is funded by the DFG Priority Programme 2392 Visual Communication (ViCom, 2022-2025), Frankfurt am Main, Germany.

## 9. Ethical considerations

Deaf signers were consulted in the development of MSL-WN and its annotation interface. Annotation for MSL-WN was performed either by deaf signers or in consultation with them.

MSL-WN is a resource for sign linguistic research. Its license excludes commercial uses. To work towards ethical applications of sign language technologies, its dataset download page provides an advisory for users to observe the best practice recommendations by Fox et al. (2023).

## 10. Limitations

The MSL-WN is a work in progress. Individual sign entries are not guaranteed to yet have synset links for all their senses. Correspondingly, synset entries should not be expected to contain a complete lists of matching synonyms for any given language.

In addition to items verified by human annotator, MSL-WN also contains automatically detected sign-synset links for signs that are presumed to be monosemous. The dataset explicitly marks whether an entry was human- or machine-verified.

As the MSL-WN builds on the lexical inventory of several other resources, the editorial decisions of those resources can affect how signs are represented in MSL-WN, e.g. regarding how fine-grained the distinction of lexical variants is.

## 11. Bibliographical References

- Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kiki Vasilaki, Anna Vacalopoulou, Theodoros Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, and Eleni Efthimiou. 2022. [Introducing sign languages to a multilingual wordnet: Bootstrapping corpora and lexical resources of Greek Sign Language and German Sign Language](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 9–15, Marseille, France. European Language Resources Association (ELRA).
- Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, Eleni Efthimiou, Neil Fox, Onno Crasborn, Lianne Westenberg, Sarah Ebling, and Laure Wawrinka. 2023. [Interlingual index for the project's core sign languages](#). Project Note D6.4, EASIER Consortium, Hamburg, Germany.
- Sam Bigeard, Marc Schulder, Maria Kopf, Thomas Hanke, Kiriaki Vasilaki, Anna Vacalopoulou, Theodor Goulas, Athanasia-Lida Dimou, Stavroula-Evita Fotinea, Eleni Efthimiou, Neil Fox, Onno Crasborn, Lianne Westenberg, Sarah Ebling, Laure Wawrinka, Johanna Mesch, Thomas Björkstrand, Anna Kuder, and Joanna Wójcicka. 2024. [Extended interlingual index for the project's core sign languages and languages covered in WP9](#). Project Note D6.5, EASIER Consortium, Hamburg, Germany.
- Francis Bond and Kyonghee Paik. 2012. [A survey of WordNets and their licenses](#). In *Proceedings of the 6th Global WordNet Conference*, pages 64–71, Matsue, Japan.
- Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. [CIL: the Collaborative Interlingual Index](#). In *Proceedings of the Eighth Global WordNet Conference*, Bucharest, Romania. University of Iasi.
- Sonja E. Bosch and Marissa Griesel. 2017. [Strategies for building wordnets for under-resourced languages: The case of African languages](#). *Literator*, 38(1):1–12.
- Alessandra Corda, Vincenzo Lo Cascio, and Massimiliano Pipolo. 1998. [Automatic reversal of a bilingual dictionary: Implications for lexicographic work](#). In *Proceedings of the 8th EU-RALEX International Congress*, pages 433–443.
- Thierry Declerck and Sussi Olsen. 2023. [Linked open data compliant representation of the interlinking of Nordic wordnets and sign language data](#). In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 62–69, Tórshavn, the Faroe Islands. Association for Computational Linguistics.
- Dicta-Sign Consortium. 2012. [Sign-Wiki prototype](#). Project deliverable D6.7, Dicta-Sign Consortium.
- Sarah Ebling, Katja Tissi, and Martin Volk. 2012. [Semi-automatic annotation of semantic relations in a Swiss German Sign Language lexicon](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 31–36, Istanbul, Turkey. European Language Resources Association (ELRA).
- Eleni Efthimiou, Stavroula-Evita Fotinea, Athanasia-Lida Dimou, Theodoros Goulas, Panagiotis Karioris, Kiki Vasilaki, Anna Vacalopoulou, and Michalis Pissaris. 2016. [From a sign lexical database to an SL golden corpus – the POLYTROPON SL resource](#). In *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 63–68, Portorož, Slovenia. European Language Resources Association (ELRA).
- Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. [Sign language technologies and resources of the Dicta-Sign project](#). In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 37–44, Istanbul, Turkey. European Language Resources Association (ELRA).
- Christiane Fellbaum, editor. 1998. [WordNet: An Electronic Lexical Database](#). The MIT Press.
- Neil Fox, Bencie Woll, and Kearsy Cormier. 2023. [Best practices for sign language technology research](#). *Universal Access in the Information Society*.
- Birgit Hamp and Helmut Feldweg. 1997. [GermaNet - a lexical-semantic net for German](#). In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain. Association for Computational Linguistics.



- Reiner Konrad, Thomas Hanke, Gabriele Langer, Dolly Blanck, Julian Bleicken, Ilona Hofmann, Olga Jeziorski, Lutz König, Susanne König, Rie Nishio, Anja Regen, Uta Salden, Sven Wagner, Satu Worseck, Oliver Böse, Elena Jahn, and Marc Schulder. 2020. *MEINE DGS – annotiert. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release / MY DGS – annotated. Public Corpus of German Sign Language, 3rd release*. Universität Hamburg. PID <https://doi.org/10.25592/dgs.corpus-3.0>.
- Maria Kopf, Marc Schulder, and Thomas Hanke. 2022a. *The Sign Language Dataset Compendium: Creating an overview of digital linguistic resources*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 102–109, Marseille, France. European Language Resources Association (ELRA).
- Maria Kopf, Marc Schulder, Thomas Hanke, and Sam Bigeard. 2022b. *Specification for the harmonization of sign language annotations*. Project Note D6.2, EASIER Consortium, Hamburg, Germany.
- Anna Kuder, Joanna Wójcicka, Piotr Mostowski, and Paweł Rutkowski. 2022. *Open repository of the Polish Sign Language Corpus: Publication project of the Polish Sign Language Corpus*. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 118–123, Marseille, France. European Language Resources Association (ELRA).
- Joanna Łacheta, Małgorzata Czajkowska-Kisil, Jadwiga Linde-Usiekniewicz, and Paweł Rutkowski. 2016. *Korpusowy słownik polskiego języka migowego/Corpus-based Dictionary of Polish Sign Language*. Faculty of Polish Studies, University of Warsaw.
- Khang Nhut Lam and Jugal Kalita. 2013. *Creating reverse bilingual dictionaries*. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 524–528, Atlanta, Georgia. Association for Computational Linguistics.
- Gabriele Langer and Marc Schulder. 2020. *Collocations in sign language lexicography: Towards semantic abstractions for word sense discrimination*. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 127–134, Marseille, France. European Language Resources Association (ELRA).
- Jurie Le Roux, Koliswa Moropa, Sonja Bosch, and Christiane Fellbaum. 2008. *Introducing the African languages wordnet*. In *Proceedings of The Fourth Global WordNet Conference*, pages 269–280, Szeged, Hungary. University of Szeged, Department of Informatics.
- Colin P. Lualdi, Elaine Wright, Jack Hudson, Naomi K. Caselli, and Christiane Fellbaum. 2021. *Implementing ASLNet v1.0: Progress and plans*. In *Proceedings of the 11th Global Wordnet Conference*, pages 63–72, Potchefstroom, South Africa. South African Centre for Digital Language Resources (SADiLaR).
- Willy Martin. 2013. *Reversal of bilingual dictionaries*. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert E. Wiegand, editors, *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, volume 5/4 of *Handbooks of Linguistics and Communication Science*, pages 1445–1455. De Gruyter Mouton, Berlin.
- Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert, and Eva Safar. 2012. *Dicta-Sign – building a multilingual sign language corpus*. In *Proceedings of the LREC2012 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*, pages 117–122, Istanbul, Turkey. European Language Resources Association (ELRA).
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. *Introduction to WordNet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244.
- Bolette S. Pedersen, Sanni Nimb, Jørg Asmussen, N. H. Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. *DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary*. *Language Resources and Evaluation*, 43:269–299.
- Siegmund Prillwitz, Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, and Arvid Schwarz. 2008. *DGS Corpus project – development of a corpus based electronic dictionary German Sign Language / German*. In *Proceedings of the LREC2008 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign*

- Language Corpora*, pages 159–164, Marrakech, Morocco. European Language Resources Association (ELRA).
- Umar Shoaib, Nadeem Ahmad, Paolo Prinetto, and Gabriele Tiotto. 2014. [Integrating Multi-WordNet with Italian Sign Language lexical resources](#). *Expert Systems with Applications*, 41(5):2300–2308.
- Svenskt teckenspråkslexikon. 2024. [Swedish Sign Language Dictionary online](#). Department of Linguistics, Stockholm University.
- Thomas Troelsgård and Jette Kristoffersen. 2018a. [Improving lemmatisation consistency without a phonological description. The Danish Sign Language corpus and dictionary project](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 195–198, Miyazaki, Japan. European Language Resources Association (ELRA).
- Thomas Troelsgård and Jette Hedegaard Kristoffersen. 2018b. [Building a sign language corpus – problems and challenges: The Danish Sign Language Corpus and Dictionary](#). Abstract published at the XVIII EURALEX International Congress Lexicography in Global Contexts.
- Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. [BalkaNet: Aims, methods, results and perspectives. A general overview](#). *Romanian Journal of Information Science and Technology*, 7(1–2):9–43.
- Anna Vacalopoulou, Eleni Efthimiou, Stavroula-Evita Fotinea, Theodoros Goulas, Athanasia-Lida Dimou, and Kiki Vasilaki. 2022. [Organizing a bilingual lexicographic database with the use of WordNet](#). In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany*, pages 357–366. IDS-Verlag.
- Piek Vossen, editor. 1998. [EuroWordNet: A multilingual database with lexical semantic networks](#). Springer Netherlands, Dordrecht.
- Piek Vossen, Francis Bond, and John McCrae. 2016. [Toward a truly multilingual GlobalWordnet grid](#). In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 424–431, Bucharest, Romania. Global Wordnet Association.

# Facial Expressions for Sign Language Synthesis using FACSHuman and AZee

Paritosh Sharma , Camille Challant , Michael Filhol 

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique,  
91400, Orsay, France

{paritosh.sharma, camille.challant}@universite-paris-saclay.fr, michael.filhol@cnrs.fr

## Abstract

This paper presents an approach to synthesising facial expressions on signing avatars. We implement those generated by a recently proposed set of rules formalised in the AZee framework for French Sign Language. Our methodology combines computer vision, linguistic insights, and morph target animation to address the challenges posed by the synthesis of nuanced facial expressions, which are pivotal for conveying emotions and grammatical cues in Sign Language. By implementing a set of universally applicable morphs and incorporating these advancements into our animation system, we aim to improve the realism and expressiveness of signing avatars. Our findings suggest an enhancement in the synthesis of non-manual signals, which extends to multiple avatars. This work opens new avenues for future research, including the exploration of more sophisticated facial modelling techniques and the potential integration of facial motion capture data to refine the animation of facial expressions further.

**Keywords:** Sign Language, Avatars, Facial expressions, AZee

## 1. Introduction

Signing avatars represent a crucial development in facilitating accessible communication for the Deaf and hard of hearing communities, enabling the visualization of Sign Language (SL) through computer-generated figures. The AZee model is instrumental in this, allowing for the synthesis of detailed multi-track animation timelines that specify the entirety of an utterance for rendering, thus enabling the creation of new SL content without the need for pre-existing animations.

Facial expressions, essential for conveying nuanced meanings in SL, pose significant challenges in the synthesis process for signing avatars. These non-manual features not only add depth and emotion to the communication but are also key in identifying meaning. The integration of facial expressions into signing avatars requires sophisticated modeling to accurately capture the wide range of emotions and grammatical cues that are communicated through subtle facial movements. This complexity makes the synthesis of facial expressions a vital area of focus to enhance the realism and effectiveness of signing avatars, ensuring they can serve as true representatives of SL communication.

This paper introduces an approach to formalizing the modeling and synthesis of facial expressions. We propose a methodology that combines computer vision, linguistic intervention, and morph target animation to improve the expressiveness and realism of signing avatars. This methodology integrates these advancements into our animation system to enhance the synthesis of non-manual signals based on a recent corpus (Challant and Filhol, 2024).

The paper is structured into sections discussing

the background research on facial expressions in SL, the methodology for creating and implementing these expressions, the results of applying this methodology, and concludes with the key findings, implications, and potential future research directions.

## 2. Background Research

Although this has not always been recognised, we now know that the use of non-manual articulators in SLs is essential: it conveys linguistic information as much as hands activity (Pfau and Quer, 2010; Crasborn, 2006; Liddell, 2003). In this paper, we only focus on facial expressions: movements of the lips, eyebrows, cheeks and the tongue. The key role of facial expressions in SLs can be clearly seen when animating avatars: the presence of facial expressions on a virtual signer considerably helps Deaf people to better understand the generated discourse (Huenerfauth et al., 2011).

In linguistic studies, face articulators are most of the time studied separately: we can find studies on eyebrows (Kimmelman et al., 2020; De Vos et al., 2009) or on mouth gestures (Lewin and Schembri, 2011). We do not account for the particular case of mouthing, which consists in articulating lips following words from a spoken language. Indeed, the phenomenon is not observed on all signers, so we decided not to give it priority in our work.

We can also note that a facial articulator is often linked to a particular grammatical phenomenon such as questions (Schalber, 2006), conditional clauses (Reilly et al., 1990) or negation (Zeshan, 2004) and recognised as belonging to a defined linguistic level: phonological, lexical or syntactic.

Nevertheless, considering articulators together (rather than separately as in traditional approaches) seems relevant. The meaning conveyed by a set of articulators is not the same as that carried by an articulator studied on its own.

We are thus interested in AZee, a formal model which allows a wholistic approach of facial expressions (Filhol, 2021; Filhol et al., 2014). Indeed, the AZee approach is based on the notion of production rule, which associates a meaning to a set of observable forms. These can be movements of the hands, arms, chest, or any part of the face: there is no hierarchy between all these articulators.

A study has just been published on facial expressions in AZee (Challant and Filhol, 2024), based on a corpus called *40 brèves* (Filhol and Tannier, 2014). It consists of 40 news items in written French, each translated into French Sign Language by three deaf translators, for a total of one hour of SL. A new set of 22 AZee production rules producing facial expressions was found (for instance *big-threatening*, *closer-look* or *with-surprise*). This covers all expressions of the corpus, which to us constitutes a substantial subset to start with for LSF animation.

While the meanings are clearly identified for the rules concerning facial expressions, a problem is that the forms have only been approximately described or captured with still shots of signers producing them. It is now necessary to describe the forms of these facial expressions more precisely in order to animate them on virtual signers.

The methods for synthesizing facial expressions in SL animations encompass a variety of techniques. These methods include manual animation based on linguistic insights, automated techniques using motion capture data, and computer vision approaches for feature extraction. Sims (2000) offer unique approaches with varying degrees of success in capturing and animating nuanced facial expressions critical to SL communication. Kennaway et al. (2007) create blend shapes for the face which map to HamNoSys. However, they group various facial parts such as eyebrows, eyelids, and nose in same tier which complicates the modeling of facial expressions where these parts of the face are not moving in parallel and could pose restrictions in co-occurring facial expressions that share some of the parts of the face. Gibet et al. (2011) utilizes motion capture for more naturalistic expressions, facing challenges in data capture and representation granularity. A set of blend shapes were rigged on the Paula avatar (McDonald et al., 2022), which can also directly map to some AZee blendshapes. However, a bigger, more comprehensive mapping is still missing.

The FACS (Ekman and Friesen, 1978) breaks down facial expressions into individual components

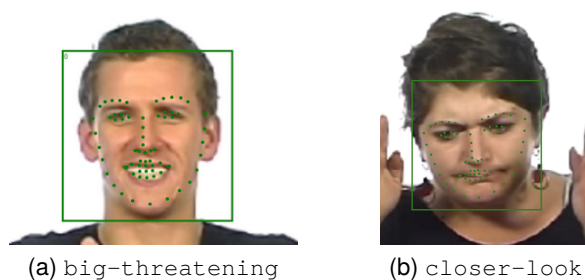


Figure 1: AU detection for two productions rules

called action units (AUs), each of which corresponds to the principle muscles responsible for that movement. FACS is used in various fields, including psychology, cognitive science, and animation, to analyze and understand emotions, intentions, and reactions through facial expressions. Gilbert et al. (2021) developed a set of blendshapes which map directly to a subset of the FACS based on a template mesh. Thus, defining our facial expressions in terms of FACS AUs and mapping the FACS Human blendshapes as AZee morphs would allow us to create a comprehensive set of facial expressions for any avatar which is based on this template mesh (Sharma and Filhol, 2023a).

### 3. Methodology

#### 3.1. Modeling

To begin with, the first step was to model the 22 facial expressions AZee production rules using the software MakeHuman. For this, we used the FACS Human plugin, which allows to model a human face thanks to different sliders, which are divided as follows:

- upper face (movements of the eyebrows, the lids and the cheeks);
- lower face (movements of the nose and the lips);
- head position;
- eye positions;
- lip parting and jaw opening;
- miscellaneous (e.g. cheek puff, tongue out, movements of the nostrils or the pupils).

Within each of these categories, there are different AUs for which the cursor can be placed between 0 (rest) and 100 (extreme position for this AU).

We worked with pictures extracted from the *40 brèves* corpus, which allowed us to find the new AZee facial expressions production rules. These





Figure 2: `big-threatening` based on a motion template (higher acceleration for jaw-drop)

pictures are easier to use when we try to model the face on an avatar than videos.

To start with, and to avoid modeling expressions from scratch manually, we tried to use automatic detection with FaceTorch (Figure 1), an AU detector based on work by Luo et al. (2022).

The detector models AU relationships and deep learns a unique graph to explicitly describe the relationship between each pair of AUs of the target face and thus detects compositant Facial AUs from single RGB images. We realised that when AUs were detected, they were most of the time correct and gave us good clues to create the blendshapes for the face. But some activations were missing and the method is anyway constrained by the lack of AU intensity specification. Thus, linguist intervention was important at every iteration during the modeling process. For example, for the rule `big-threatening`, Luo et al. (2022) detects the following AUs (see figure 1): Brow Lowerer, Cheek Raiser, Lid Tightener, Upper Lip Raiser, Lip Corner Puller, Lips Part. When we tried to model the rule using FACSHuman, we used more AUs than what was detected initially : Inner Brow Raise, Outer Brow Raise (Left and Right), Eye Closure, Nose Wrinkle, Sharp Lip Puller (Left and Right), Dimpler, Lip Stretch (Left and Right), Lip Funneler (Bottom Lip and Both Lip), Lips Suck (Lower lip), Jaw Drop Bottom Lip Down.

Most of 22 AZee rules were therefore modelled manually with FACSHuman, without using FaceTorch (Luo et al., 2022).

### 3.2. Creating Shape keys

We model all FACSHuman AUs as Blender *shape keys*, using the *target* specification to define the bending of mesh at extreme positions. For example, the target file specifies vertex adjustments for facial movements, such as “4 0.002 1” to move vertex 4 by .002 units along the Y axis (labelled “1”) for the extreme configuration.

During synthesis, these shape keys are modified as parts of *Facial Morph constraints* based on the AZee expression being synthesized. The avatar is then constrained based on these shape keys for the particular block.

### 3.3. Intermediate blocks

We extend out intermediate block generator (Sharma and Filhol, 2023b) algorithm to create interpolations for facial morphs as well. For this, we add additional *motion curves* (curves defining the displacement of vertices effected by the AU with respect to time) in the intermediate blocks based on the motion template.

This gives us a controllable motion curve profile for every AU for the facial morphs (Figure 2).

## 4. Evaluation

This section presents the evaluation of our methodology in synthesizing facial expressions for signing avatars using the AZee model. We evaluate the



Figure 3: Synthesis of `closer-look` (bottom) for male, female and neutral gender and their neutral expression for reference (top)

avatar’s ability to perform a wide range of AUs and the synthesis of the modeled facial expressions. The accompanying videos of this research can be found at <https://doi.org/10.5281/zenodo.10912305>. Video “all\_action\_units” demonstrates the full range of AUs synthesized by our avatar. Video “all\_expressions” shows all the synthesized expressions based on the French Sign Language corpus (Challant and Filhol, 2024). Figure 3 illustrates the synthesis of the expression `closer-look` across avatars of different genders, showcasing our method’s adaptability to various avatar designs. Additionally, video “big\_threatening\_hot” demonstrates the expression `big-threatening(hot())` and `hot()` alone without non-manuals illustrating the added depth and meaning when non-manual signals are incorporated.

Our observations confirm that the avatars can perform a substantial range of recognizable expressions. The ability to apply these expressions across different avatars with no limitations underscores the universality of our methodology. However, we feel that the current model can be further improved to capture more nuanced expressions. We have indeed encountered a few limitations when we tried to model the different productions rules. All the limits are detailed in Table 1.

## 5. Conclusion

Integrating FACSHuman and AZee, our methodology overcomes challenges in facial expression synthesis for signing avatars, applicable on a series of avatars based on the same template, and descriptively rich expressions, enhancing both realism and communicative clarity. Our approach represents a significant advancement in the animation of facial expressions for sign languages, utilizing state-of-the-art methods in sign language representation (AZee) and animation (building face shapes from recognition). By combining these techniques, we ensure that facial animations not only accurately represent the intended expressions but also maintain fidelity to the intricacies of sign language communication, thereby enhancing the overall user experience and effectiveness of sign language avatars.

The natural next step now is to include these expressions on sample utterances (e.g. AZee sub-expressions from the attested data), and run them by LSF users for a more systematic evaluation. This approach will facilitate a deeper understanding of how well the expressions are understood and received within the LSF community and also give us insights on potential improvements (range of action units, acceleration information, etcetera).

Another potential improvement on our system could be a better facial model. Recent works such as Li et al. (2017) and Qin et al. (2023) use similar

Expression	Limitations
almost-reaching	Mouth modeling unconvincing.
continuously	"Pffff" air and cheek puff difficult, neutral eyebrows.
do-you-realise	Thick eyebrow issue.
it-is-a-shame	Mouth expression not quite real.
most-probably	Less visible teeth preferred, thick eyebrow issue.
much-almost-too-much	Frowning eyebrows and lack of eye wrinkles not convincing.
nothing-sticks-out	Tucked lips difficult to model.
something-sticks-out	Interpreted as confusion, mouth modeling limitation.
trouble-disturbance	Frowning eyebrows difficult, mouth "rising" hard to model, result not convincing.
uneasy-awkward	Tongue tip out with slightly open mouth hard to model, unconvincing.
with-chaos	Single cheek blow/puff and alternating eye blinks hard without animation.
with-no-precision	Upper lip over lower and mouth near nose unmodellable.
with-surprise	Cannot lower lower eyelid fully, thick eyebrow issue.
with-uncertainty	Appears sadder than uncertain, thick eyebrow issue.
with-worry	Lack of wrinkles around nose/forehead.

Table 1: Limitations for each facial expression rule.



Figure 4: Better facial expressions achieved using FLAME (Li et al., 2017)

philosophy of using a template mesh but generate better facial expressions since their models also account for other parameters such as stretching of skin and underlying muscles. This is demonstrated in figure 4 where the expression was generated manually using the first 100 principle components of the FLAME model. However, this generation can be automated using a flame-compatible recognition technique such as EMOCA Daněček et al. (2022). Another potential area of improvement could be the automatic creation of motion templates from retargeted facial motion capture data, thus adding much more detail to the interpolations.

The potential impact of having facial expressions with signing avatars is substantial. It enhances the

capabilities of signing avatars making them much more expressive and realistic and opens new pathways for research and development in SL synthesis.

## 6. Acknowledgements

This work has been funded by the Bpifrance investment "Structuring Projects for Competitiveness" (PSPC), as part of the Serveur Gestuel project (IVès et 4Dviews Companies, LISN — University Paris-Saclay, and Gipsa-Lab — Grenoble Alpes University).

## 7. Bibliographical References

- Anastasia Bauer and Masha Kyuseva. 2022. [New insights into mouthings: Evidence from a corpus-based study of russian sign language](#). *Frontiers in Psychology*, 12.
- Camille Challant and Michael Filhol. 2024. Extending AZee with Non-manual Gesture Rules for French Sign Language. In *Proceedings of the 14th Language Resources and Evaluation Conference (LREC)*, Torino, Italy.
- Onno Crasborn. 2006. [Nonmanual structures in sign language](#). In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 668–672. Elsevier, Oxford.

- Radek Daněček, Michael J Black, and Timo Bolkart. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322.
- Connie De Vos, Els Van Der Kooij, and Onno Crasborn. 2009. Mixed signals: Combining linguistic and affective functions of eyebrows in questions in sign language of the netherlands. *Language and Speech*, 52(2–3):315–339.
- Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.
- Michael Filhol. 2021. *Modélisation, traitement automatique et outillage logiciel des langues des signes*. Habilitation à diriger des recherches, Université Paris-Saclay.
- Michael Filhol, Mohamed Hadjadj, and Annick Choisier. 2014. Non-Manual Features: The Right to Indifference. In *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.
- Michael Filhol and Xavier Tannier. 2014. Construction of a French–Lsf Corpus. In *Building and Using Comparable Corpora, Language Resource and Evaluation Conference (LREC)*, page 4, Reykjavik, Iceland.
- Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The SignCom system for data-driven animation of interactive virtual signers: Methodology and Evaluation. *ACM Trans. Interact. Intell. Syst.*, 1(1).
- Michaël Gilbert, Samuel Demarchi, and Isabel Urdapilleta. 2021. FACSHuman, a software program for creating experimental material by modeling 3D facial expressions. *Behavior Research Methods*, 53(5):2252–2272.
- Thomas Hanke. HamNoSys – Representing Sign Language Data in Language Resources and Language Processing Contexts.
- Matt Huenerfauth, Pengfei Lu, and Andrew Rosenberg. 2011. Evaluating importance of facial expression in American Sign Language and pidgin signed English animations. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, page 99–106, Dundee Scotland, UK. ACM.
- Hernisa Kacorri. 2015. *TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation*. CUNY Graduate Center.
- J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood. 2007. Providing signed content on the internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact.*, 14(3):15–es.
- Vadim Kimmelman, Alfarabi Imashev, Medet Mukushev, and Anara Sandygulova. 2020. Eyebrow position in grammatical and emotional expressions in Kazakh-Russian Sign Language: A quantitative study. *PLOS ONE*, 15(6):e0233731.
- Donna Lewin and Adam C. Schembri. 2011. Mouth gestures in British Sign Language: A case study of tongue protrusion in BSL narratives. *Sign Language & Linguistics*, 14(1):94–114.
- Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17.
- Scott K. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge University Press.
- Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning multi-dimensional edge feature-based au relation graph for facial Action Unit recognition. *arXiv preprint arXiv:2205.01782*.
- John McDonald, Ronan Johnson, and Rosalee Wolfe. 2022. A Novel Approach to Managing Lower Face Complexity in Signing Avatars. In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 67–72, Marseille, France. European Language Resources Association.
- Roland Pfau and Josep Quer. 2010. *Nonmanuals: their Grammatical and Prosodic Roles*, pages 381–402. Cambridge University Press.
- Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. 2023. Neural Face Rigging for Animating and Retargeting Facial Meshes in the Wild. *arXiv preprint arXiv:2305.08296*.
- Judy Snitzer Reilly, Marina McIntire, and Ursula Bellugi. 1990. The acquisition of conditionals in American Sign Language: Grammaticized facial expressions. *Applied Psycholinguistics*, 11(4):369–392.
- Katharina Schalber. 2006. What is the chin doing?: An analysis of interrogatives in Austrian Sign Language. *Sign Language & Linguistics*, 9(1-2):133–150.



- Paritosh Sharma and Michael Filhol. 2023a. [Extending Morphs in AZee Using Pose Space Deformations](#). In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.
- Paritosh Sharma and Michael Filhol. 2023b. [Intermediate block generation for multi-track sign language synthesis](#). In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '23*, New York, NY, USA. Association for Computing Machinery.
- Ed Sims. 2000. [Virtual communicator characters](#). *SIGGRAPH Comput. Graph.*, 34(2):44.
- Ulrike Zeshan. 2004. [Hand, head, and face: Negative constructions in Sign Languages](#). *Linguistic Typology*, 8(1):1–58.

# Eye Blink Detection in Sign Language Data Using CNNs and Rule-Based Methods

Margaux Susman , Vadim Kimmelman 

University of Bergen  
Sydnesplassen 7, 5007 Bergen, Norway  
{margaux.susman, vadim.kimmelman}@uib.no

## Abstract

Eye blinks are used in a variety of sign languages as prosodic boundary markers. However, no cross-linguistic quantitative research on eye blinks exists. In order to facilitate such research in future, we develop and test different methods of automatic eyeblink identification, based on a linguistic definition of blinks, and in a dataset of a natural sign language (French Sign Language). We compare two main approaches to eye openness detection: calculating the Eye Aspect Ratio using MediaPipe, and training CNNs to detect openness directly based on images from the video recordings. For the CNN method, we train different models (with different numbers of signers in the training data, different frame crops and different numbers of epochs). We then combine the openness degree detection with a separate rule-based component in order to determine boundaries of blink events. We demonstrate that both methods perform relatively well, and discuss the practical implications of the methods.

**Keywords:** Blink detection, French Sign Language, Machine Learning

## 1. Introduction

Eye blinks are a natural physiological phenomenon which is independent of speech and language production, but which is also involved in language production in various ways. Crucially, in sign languages, eye blinks have been shown to serve as boundary prosodic markers (Sze, 2008). Some studies indicate that eye blinks are co-occurring with prosodic units, and that different sign languages employ eye blinks differently, that is, at different levels at the prosodic structure (ibid.). However, currently, no large quantitative research on eye blinks in sign languages exists, neither for specific sign languages, nor for the purposes of cross-linguistic comparison.

In order to make such research possible, it is necessary to have a reliable method of automatically identifying and annotating blinks in video recordings of signers communicating in a signed language. Due to recent advances in Computer Vision (CV) and Deep Learning, it is now possible to attempt this. In fact, the blink detection task has been pursued in many studies (Dewi et al., 2022; Fodor et al., 2023; Hong et al., 2024), but not specifically using sign language data or with sign languages in mind. In addition, the definition of blinks in the blink detection literature is quite different from the linguistic understanding of eye blinks in sign linguistics.

In this paper, we report a study in which we implemented and tested two proofs of concept of eye blink detection in a corpus of French Sign Language (LSF). We tested two main methods: a combination of a newly trained CNN associated with two different rule-based blink identification methods, and a combination of an existing CV solution, using Me-

diaPipe (Grishchenko and Bazaressvky, 2020), with a simple eye aspect ratio (EAR) calculation, also followed by the rule-based algorithms. We specifically test how the number of signers in the dataset, as well as other specific methodological decisions influence the success of eye blink identification.

## 2. Background

### 2.1. Eyeblinks in communication

#### 2.1.1. Physiology of blinks

In physiology, blinks have been defined as having three phases, that is a closing phase, a closed phase and a reopening phase. They have also been differentiated from closures as they last longer and do not carry the same meanings and functions as blinks do in communication. Stern and Skelly (1984) note that “for blinks, the time from initiation of lid movement to full eye closure is short, [...] less than 150ms, whereas for non-blink closures, the time taken to close the eyes is [...] generally greater than 250ms and frequently extends over a period of seconds.” Blinks may exhibit an incomplete closure of the lids (Sforza et al., 2008).

As was noted by Ponder and Kennedy (1927) but also Hall (1945), Karson et al. (1981) and Bentivoglio et al. (1997), blink rates in conversations is higher than while resting or reading. Hall (1945) reports an average blink rate of 25.4 blinks per minute in conversation against an average blink rate of 3.29 blinks per minute while reading. Similar average blink rates while speaking are reported by Karson et al. (1981) and Bentivoglio et al. (1997). Hömke et al. (2017) suggested that blink events occur at

turn taking points in conversations.

Finally, [Descroix et al. \(2022\)](#) recently investigated blinking in spoken communication. They found that in addressees, the blinking rate depends on the degree of interest in the communicated information; when presented with an interesting message, the blink rate of the addressee increases. On the other hand, the blink rate of the speaker is said to be higher than while being silent and alone, regardless of the interest to the shared information. The authors note that these findings give evidence to the “interactive communication function of Spontaneous Eye Blinks”.

### 2.1.2. Blinks in Sign Languages

Several researchers working on a variety of signed languages have argued that eye blinks have a linguistic function ([Baker and Padden, 1978](#); [Wilbur, 1994](#); [Sze, 2008](#)). For example, [Wilbur \(1994\)](#) argued that some eye blinks<sup>1</sup> in American Sign Language (ASL) occur at the end of Intonation Phrases, and thus serve as prosodic boundary markers, while other blinks occur on lexical signs and have lexical or emphatic functions.

[Sze \(2008\)](#) investigates eye blinks in Hong Kong Sign Language (HKSL). She finds both similarities and differences in the functioning of blinks in this language. Specifically, for prosodically aligned blinks (“boundary-sensitive” in her terminology), she argues that they do not necessarily align with Intonational Phrases. According to her, they occur at the potential Intonational Phrase boundaries in 55% of the cases, while in the rest of the cases they occur at boundaries of other and typically smaller prosodic/grammatical units. In addition, she demonstrates that eye gaze change and head movement can lead to the use of blinks, even in the absence of linguistic boundaries.

The issue of classifying the functions/types of blinks is thus very complicated. The classifications in [Wilbur \(1994\)](#) and [Sze \(2008\)](#) differ in the level of detail, and these authors would classify some of the blinks quite differently. In a recent study of LSF ([Chételat-Pelé, 2010](#)), yet another classification was applied.

To summarize, a few studies have shown that blinks have important linguistic functions in sign languages. Note however, the following limitations. First, only a handful of sign languages have been studied so far. Second, while all the researchers note that blinks often align with (prosodic) boundaries, more specific functions attributed to blinks vary between the different studies, and thus a comparison is not possible. Third, the datasets analyzed in the existing studies are quite small. It is

<sup>1</sup>Clearly not all blinks have a linguistic function, as all the authors acknowledge, see also the previous section.

clear that much more research is necessary on this issue, including using larger datasets and analyzing blinks across multiple sign languages, multiple genres, and across individual signers. This can be achieved if automatic blink detection is available.

## 2.2. Eyeblink Detection

Sign Language Recognition (SLR) is a task at the intersection with CV and Natural Language Processing (NLP). SLR is concerned with the automatic recognition of signs and their translation into written or spoken language. Over the years, SLR methods have improved and nonmanuals started to be integrated into such recognition algorithms but as reported by [Koller \(2020\)](#), eye gaze and eye blinks have never been taken into consideration. For this reason, we turned ourselves towards blink detection algorithms. Those algorithms have mostly been implemented to solve tasks such as driver drowsiness analysis, attention level measure and eye fatigue measure ([Fodor et al., 2023](#)).

Eye blink detection methods can be divided into two categories: sensor-based methods and vision-based methods, the latter having become more popular in recent years ([Hong et al., 2024](#)).

[Soukupová and Cech \(2016\)](#) introduced the Eye Aspect Ratio (EAR) as a measure of eye openness. They report that the EAR is an estimation of the degree of openness of the eye. The EAR is the calculation of the distances between the lower and upper lids (with two computations per eye) and of the distance between the left and right corners of each eye. The equation of the EAR measurement is presented in (1) and the placement of the points  $P$  is shown in figure 1.  $P_n$  are landmarks locations represented in 2D.  $P_1$  is the landmark denoting the outside part of the eye,  $P_4$  denotes the inside part of the eye while  $P_2$  and  $P_3$  both denote point on the upper lid and  $P_5$  and  $P_6$  denote point on the lower eyelid.

$$EAR = \frac{\|P_2 - P_6\| + \|P_3 - P_5\|}{2 \|P_1 - P_4\|} \quad (1)$$

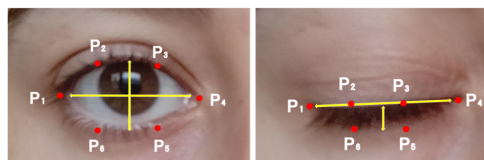


Figure 1: Eye landmarks position for the EAR calculation with open eye and with closed eye.

This EAR calculation has been widely used ([Ibrahim et al., 2021](#); [Dewi et al., 2022](#); [Phuong et al., 2022](#)).

Recent blink detection studies include tasks addressing issues such as computation cost (Ibrahim et al., 2021), luminosity changes (Dewi et al., 2022), head movements (Hong et al., 2024). Most methods start by extracting the eye region using face detection and facial landmark detection methods or apply the EAR calculation combined with deep learning architectures (Hong et al., 2024). Another issue that we address in this present study is the lack of consideration for blinks with incomplete lid closure, which are considered blinks (Sforza et al., 2008) and which we encounter in our data.

### 3. Methods

In the present research, we aim at detecting blinks automatically in sign language data in order to facilitate further linguistic analysis of blinks. We define a blink as a rapid closure and reopening of the lids, limited in duration and which can exhibit an incomplete closure. Would using a machine learning (ML)-based algorithm combined with a rule-based model improve the detection of blinks so defined?

Previous eye blink detection algorithms have failed to encompass the incompleteness of lids closure and the restriction on their duration. To address those shortcomings, we present a proof of concept that combines a ML-based classifier that determines the degree of openness of an eye (open, in-between, closed) with a rule-based model taking a set window of frames as input to determine whether a blink is occurring or not.

#### 3.1. The Dataset

We work using sign language data. We use a subpart of the Dicta-Sign corpus (Matthes et al., 2010), namely the Dicta-Sign-LSF-v2 (Belissen et al., 2020). The dataset contains video recordings of discussion about European travel. The content of those recordings was loosely elicited. In the LSF subpart, nine dyads of signers are conversing. Each of the 18 signers performed between 3 and 9 tasks. Videos and the partial annotations of the data are available [online](#). The annotated data includes glosses for the right and left hands as well as glosses for signs articulated with both hands. In the annotated files, the annotation of a gloss is represented by an ID which is linked to a gloss in a separate document. A subset of videos is annotated for loose translations of the signed utterances. For this study, we select a subset of the annotated data.

We use data from 5 different participants. Information about the signers is available in Table 1.

Signer	G	Age	Learn LSF	Deaf fam.
A11	F	28	biling. school	no
B15	M	38	prim. school	yes
B14	F	28	kindergar.	yes
A9	F	28	birth	yes
B5	F	28	birth	yes

Table 1: Participants' metadata

#### 3.2. Annotation

We annotated the blink occurrences using ELAN 6.2 software program (Sloetjes and Wittenburg, 2008). Videos were captured at 25fps. The shortest video consists of 5500 frames while the longest video contained over 16500 frames. The .csv files containing the original annotations of the corpus (Belissen et al., 2020) are transformed so that the frames are converted into time intervals using a Python script. A second Python script is used to connect the ID of the annotation to the gloss of its sign. As part of the current project, a total of 26 videos were annotated, that is a total of 2 hours and 59 minutes and 4342 blinks, giving an average of 24 blinks per minute. For the experiments conducted in this paper, we selected 9 videos, that is 60 minutes and 36 seconds and a total of 1565 annotated blinks, giving an average of 26 blinks per minute. 4 of the videos are used for the training of the various ML models while we apply the blink detection algorithm to the other 5.

Sze (2008) divides blinks into three phases, specifically the closing of the lid, the eyes closed and finally its reopening. On the other hand, (Chételat-Pelé, 2010) divides the blink into two phases, that is the closing of the lid and its reopening. She adds that the full closure of the lid should not exceed 40 milliseconds limit above which we observe a closed eye and not a simple blink anymore.

In this study, one annotation for a blink includes all three phases, and its duration covers the three phases, as motivated by the definition of blinks given by physiologists. In cases in which the lids reopen to squinted eyes for example, we stop the annotation of the blink at the frame where the lid is not opening further, while in regular cases, we stop the annotation when the lid excursion is back to what it was prior to the blink event. The annotation was conducted by the first author, with discussion of specific cases with the second author.

In the data of Chételat-Pelé (2010), the shortest recorded blink lasts 160 milliseconds, while the longest doesn't exceed 380 milliseconds. We obtain similar results with a mean blink duration across all signers of 230 milliseconds over our whole dataset and 233 milliseconds in the selection of 9 videos as shown in Table 2.



Video	Vid. duration	Blinks	Av. blink duration	Shortest blink	Longest blink
S2T1B15	11:05:000	229	0.215s	0.12s	0.60s
S9T1B5	10:35.240	282	0.204s	0.08s	0.48s
S5T9A9	05:21.823	206	0.236s	0.11s	0.39s
S4T4B14	06:14.560	204	0.250s	0.09s	0.63s
S2T2B15	03:51.000	100	0.240s	0.13s	0.68s
S9T2B5	04:07.520	156	0.243s	0.11s	0.61s
S5T3A9	05:47.680	159	0.239s	0.09s	0.50s
S4T7B14	09:41.040	166	0.219s	0.07s	0.55s
S2T3A11	04:28.000	63	0.259	0.13s	0.73s

Table 2: Blink annotation statistics

### 3.3. Automatic Blink Detection

In the field of automatic blink detection, blink events have rarely been defined and when it was done, the issue is described as a state of openness task rather than a blink detection task. Zeng et al. (2023) claim creating an eye blink detection model but compare their work to Phuong et al. (2022) who use “eye blink detection” in the title of their paper but keep noting that they are “propos[ing] a technique to detect the open/closed state of the eyes”. Dewi et al. (2022) write: “We can assume that the eye is closed/blinked when: (1) Eyeball is not visible, (2) eyelid is closed, (3) the upper and lower eyelids are connected.” Two problems arise from such a description of blinks: this definition (1) does not account for incomplete blinks (2) nor for closures which last typically longer than blinks.

Making the task a binary one, with *open* and *closed* classes is overseeing the *in-between* frames which exhibit an eye not completely closed nor completely open.

We use two methods for the detection of the eyes’ degree of openness. We use Mediapipe to detect eyes landmarks on which we use the EAR measure on one hand and, on the other hand, we train a novel ML model.

#### 3.3.1. State of Openness Detection

Before training the ML model, we create a dataset specifically for the task. We transform a subset of the annotated videos into images. We create two different crops of each frame: a face crop and an eyes crop. We use Mediapipe Face Landmarks (Grishchenko and Bazaresvsky, 2020) to determine which region of the frames needs to be cropped. Depending on the frame, the crop varies in dimension. The images are divided into an *open* and *closed* folders based on our annotations (the frames overlapping with the blink annotations are placed in the closed folder). We create a third *in-between* folder and rearrange the data across those three folders image by image. Indeed, as all three phases of a blink are annotated as one event, in the

*closed* folder, we have eyes half open. We apply this to 4 videos from 4 of our signers, namely B14, B15, A9, and B5. The *in-between* folder contains instances where the eyeball is not completely visible nor completely hidden, instances where the eye looks open but the signer keeps their head down, and instances where the eyes are hidden in cases where a sign is performed on the face. These observations reinforce the idea that a binary classification of eye openness is not ideal.

We use the EAR measurement to detect the eye openness degree. The EAR-based method includes extracting the relevant eye landmarks with Mediapipe and calculating the EAR value for each frame using the formula above. This is done in real time.

Another way of determining the eye openness degree can be done using ML techniques. We choose to use a Convolutional Neural Network (CNN) as we are working with images and CNNs are designed to treat such data. We create a CNN architecture inspired by the classic LeNet-5 architecture (Lecun et al., 1998). Our model consists of several blocks, each one includes a convolutional layer followed by a pooling layer to seize spatial correlation in the image at varying scales. The CNN ends with linear (or “fully-connected”) layers. The model for the face crops is a bit more complicated and contains an extra convolutional layer to account for the larger spatial dimensions of input images (256x256 vs. 64x128). Specifically, the face crops model is made of four convolutional layers (against three for the eyes crops model). The size of the first layer also goes up from 2080 input features for the eyes to 9216 input features for the face crops. Aside from this, the models are the same: each convolutional layer is followed by a MaxPooling layer, followed by a flattening layer and two linear layers. All layers except the last are followed by the ReLu activation function to account for non-linearity. For both models, the last layer takes 80 nodes as input and has three output features, that is one per class (open, in-between, closed). The last layer of our networks is a softmax layer that outputs a vector of proba-

bilities. We use the cross-entropy loss to calculate the distance between the probabilities given by the model and our groundtruths. Eventually, we use the Adam Optimizer to minimize complex linear functions.

### 3.3.2. Pipeline: State of Openness Detection Using Machine Learning

We create four models, each model is respectively trained on 1 signer, 2 signers, 3 signers, and 4 signers and we compare the results.

Once the data is ordered in the three folders (open, in-between, closed), we proceed and load the images. Using the PyTorch library (Paszke et al., 2019), we start developing our method. Our first step is to separate the images into a training, a validation and a test set. The preprocessing of the frames varies depending on whether those are in the training set or in the validation and test sets. We resize the frames in each sets to 256x256 for the face crops and 64x128 for the eyes crops and we convert those images into numerical values. For frames in the training set, we use the Trivial Augmentation Wide transform developed by Muller and Hutter (2021) and implemented in PyTorch. The frame distribution across our three classes is greatly unbalanced. Indeed, *open* received the vast majority of the data. If we take video S2T1 from signer B15 which we use in the training of all the models, we note that out of the 16298 frames distributed across the three categories, only 690 frames belong in the *closed* folder while the two remaining folders share the 15608 images evenly.

We recreate balance in an artificial way as we fix the number of training images on a percentage of the minority class. We fix the percentage at 70% of the minority class. For example, 70% of the 690 images mentioned earlier are used in the training set for the 1 signer model. The training set therefore contains 482 images from each of the classes. The remaining 30% of the *closed* folder are divided into two: half of the frames goes to the validation set and the other half to the test set. The rest of the frames from the two other classes are also divided into a validation and test sets.

The training set is quite small due to the under-sampling applied thus we use the virtual data augmentation method to modify the images within the training set randomly, that is, from one batch to another the images will appear differently. To this end, we use the TrivialAugment, an automatic augmentation method. The degree of transformation of an image fluctuates randomly but as noted by (Muller and Hutter, 2021), only one augmentation method is applied to the image at a time. The augmentation techniques applied to the images involve modifications of brightness, colors, contrast, blurring and sharpness along with image rotation and

image flipping transformations.

All models are trained on the eyes crops for 100 and 200 epochs and on the face crops for 100 and 200 epochs as well.

### 3.3.3. Agglomeration Over Time Using Logic-Based Rules

Once we obtained our CNN results, we create the rules which will allow making a decision as to whether or not a blink is occurring.

We use the original groundtruths (data annotated with ELAN) as .csv files, one file for one video. As a blink occurs over a set of frames, a decision is made on a window of frames representing a time interval. We split the videos into non-overlapping windows of five frames each. We implement two different rules to detect whether we observe a blink event. Those rules are the high-low-value-difference rule (R1) and the curve rule (R2). Each one will be combined with the CNN outputs on one hand and with the EAR measurement on the other hand.

The high-low-value-difference rule looks at the maximum amplitude between the values within the selected window. According to our definition of a blink, the eye should still be somewhat open at the beginning and at the end of a blink, therefore we should observe low and large values within the window of frames when a blink happens. The difference in values between the frames of a unique window should be higher than the defined threshold when there is a blink event.

As we expect the CNN and EAR values to be lower in the middle of the window of frames and higher on the outskirts of this window when a blink is occurring, we implement the curve rule. We expect the values to form a U-shaped curve. We fit a second-degree polynomial using the polynomial regression model from Scikit-Learn (Pedregosa et al., 2011). A blink occurs when the curve goes down and up steeply, and we define the steepness with a threshold.

### 3.3.4. Pipeline: Blink or Not?

We want to make a decision as to whether a blink is happening or not based on a time interval lasting longer than the duration of a single frame. Therefore, we create windows of frames. The size of the window is set at 5 when no blink occurs and follows the length of the blink otherwise. We have a large class imbalance with more intervals without blinks than with blinks thus we use the  $f1$ -score as our evaluation metric.

We compare the two rules, namely the high-low-value-difference rule with the curve rule within two methods (CNN and EAR). For the Convolutional Neural Networks, each rule is tested for the four

trained models, noting that each of these four models has been trained for 100 and 200 epochs on the eyes crops and on the face crops. We combine the EAR measurement to each of the rules as well.

We test several thresholds which differ for the CNNs and for the EAR measurement. For the CNNs, we test 8 threshold values, specifically 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. These thresholds represent the CNNs outputs probability of belonging into one of our three classes. For the EAR calculation, we test the following threshold values: 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18. Best thresholds for the EAR have been said to be contained between 0.18 and 0.20 (Soukupová and Cech, 2016) but others have criticized those thresholds and have mentioned that a greater variation can be observed (Dewi et al., 2022). The threshold represents the difference between the open and closed eyes needing to be observed for a blink to occur.

## 4. Results

### 4.1. CNN Training Results

In Table 3, we report the results obtained on the evaluation of the CNNs.

Strangely, the worst results are not obtained on the 1 signer model but rather on the 2 signers model and stay high with our lowest micro  $f_1$ -score at 80.6% and the respective macro and weighted  $f_1$ -score reaching 93.2% and 94.4% respectively.

The same way, for all eyes and face models, the best results are not exhibited for the 4 signers model but for the 3 signers model although the difference is slight. Our highest macro and weighted  $f_1$ -scores each reach 97.3% and are obtained on the face model trained for 100 epochs.

### 4.2. Blink Detection Results

After the training of the CNN, we have seen that we obtained the best evaluation results on the 3 signers model. Combined with the rules, let us see whether the 3 signers model obtains the best results. We will also look at which of the CNN or the EAR combined with the rules is best suited for our problem.

In fact, we note that the best results across four out of the five signers are obtained using the 4 signers model. The 3 signers model trained on eyes crops for 200 epochs gives the best results for the fifth signer, i.e. signer B15 with an  $f_1$ -score of 97% with 0.5 as best threshold. Results for the four other signers include  $f_1$ -scores spanning between 75% to 91.7% as can be seen in Table 4 where we report the results obtained with the four signers model and where E stands for Eyes and F for Face.

<b>1 signer</b>	<b>E 100</b>	<b>E 200</b>	<b>F 100</b>	<b>F 200</b>
test loss	0.145	0.210	0.188	0.213
test acc.	0.952	0.947	0.945	0.945
macro f1	0.848	0.835	0.840	0.833
micro f1	0.952	0.947	0.945	0.945
weighted f1	0.957	0.954	0.951	0.952
<b>2 signers</b>	<b>E 100</b>	<b>E 200</b>	<b>F 100</b>	<b>F 200</b>
test loss	0.162	0.262	0.211	0.173
test acc.	0.948	0.932	0.939	0.961
macro f1	0.840	0.806	0.819	0.869
micro f1	0.948	0.932	0.939	0.961
weighted f1	0.954	0.944	0.948	0.964
<b>3 signers</b>	<b>E 100</b>	<b>E 200</b>	<b>F 100</b>	<b>F 200</b>
test loss	<b>0.110</b>	<b>0.130</b>	<b>0.122</b>	<b>0.139</b>
test acc.	<b>0.966</b>	<b>0.968</b>	<b>0.973</b>	<b>0.970</b>
macro f1	<b>0.946</b>	<b>0.949</b>	<b>0.959</b>	<b>0.952</b>
micro f1	<b>0.966</b>	<b>0.968</b>	<b>0.973</b>	<b>0.970</b>
weighted f1	<b>0.966</b>	<b>0.969</b>	<b>0.973</b>	<b>0.971</b>
<b>4 signers</b>	<b>E 100</b>	<b>E 200</b>	<b>F 100</b>	<b>F 200</b>
test loss	0.160	0.191	0.137	0.173
test acc.	0.955	0.955	0.964	0.963
macro f1	0.937	0.937	0.951	0.951
micro f1	0.955	0.955	0.964	0.963
weighted f1	0.955	0.956	0.964	0.963

Table 3: CNNs evaluation results

For each signer, the eyes models are overall better than the face crops models. In addition, the best results are all obtained with rule 1 (R1), that is the high-low-value-difference rule, that is also true for the EAR calculations (Table 5). Concerning the EAR measurements, except for one signer, the CNN models combined with R1 gives better results than the EAR calculation combined with R1 as we see in Table 5 (where, in the parentheses of the last column, the number represents the signer model, E stands for eyes, 100 or 200 for the number of epochs the CNN has been trained and R1 stands for the high-low-value-difference rule). The difference is minimal except for signer B14 for whom we observe a 12 points difference.

Signer A11 is the only one whose data has not been used for training any of the models. In Table 6, we show the evolution of the results obtained on signer A11 across the four models. We note that for the face crops the best results are attained on the three signer model. This is in agreement with what we have seen of the evaluation of the training of the CNN models. Overall we see that the results for signer A11 are getting much better when the number of signers the CNN has been trained on increases.

We achieved the best results using the four signer CNN models combined with the high-low-value-difference rule, yet we note that the variation across signers is important and while we obtain

Signer	Eyes		Face	
	100, R1	200, R2	100, R1	200, R2
B15	0.964 [0.5]	0.969 [0.6]	0.953 [0.5]	0.953 [0.6]
B5	0.874 [0.5]	<b>0.917</b> [0.5]	0.874 [0.5]	<b>0.917</b> [0.5]
B14	<b>0.758</b> [0.9]	0.724 [0.9]	0.743 [0.7]	0.728 [0.9]
A9	0.870 [0.8]	0.822 [0.8]	<b>0.888</b> [0.7]	0.863 [0.8]
A11	<b>0.751</b> [0.8]	0.727 [0.8]	0.629 [0.5]	0.636 [0.8]

Table 4: Results of the 4 signer models

Signer	Rule 1	Rule 2	CNN best
B15	0.943	0.877	<b>0.970</b> (3, E, 200, R1)
B5	<b>0.944</b>	0.884	0.917 (4, E, 200, R1)
B14	0.638	0.650	<b>0.758</b> (4, E, 100, R1)
A9	0.874	0.806	<b>0.888</b> (4, F, 100, R1)
A11	0.723	0.650	<b>0.751</b> (4, E, 100, R1)

Table 5: Results of EAR combined with R1 and R2.

Mod.	Eyes		Face	
	100, R1	200, R2	100, R1	200, R2
1	0.611	0.661	0.594	0.617
2	0.561	0.561	0.309	0.521
3	0.705	0.681	<b>0.723</b>	<b>0.688</b>
4	<b>0.751</b>	<b>0.727</b>	0.629	0.636

Table 6: Evolution of the results across the signer models for signer A11

$f1$ -scores in the 90% for some signers, we also get  $f1$ -scores around 75% for other signers. Let us try to understand why.

When we introduced the dataset, we mentioned that the data had been loosely elicited. The signers had access to screens placed between the signers at a low height, therefore in some videos, the signers spend part of the time with their heads down, directed towards that screen. The blinks are noticeable but with difficulty, and while they have been annotated manually, the difference between the open eye and the closed one might not be enough for the models to detect it.

## 5. Discussion and Outlook

We have seen that using data from different signers in the training of the CNN allows us to obtain better results. However, we noted that the best evaluation results were obtained on the 3 signers model. We can ask ourselves whether there is a limit in terms of number of signers before the models start having less performing results. Training the CNN on more signers would allow us to test this hypothesis.

We have demonstrated that both a CNN-based approach and an EAR-based approach (which uses an existing CV solution, MediaPipe), perform the

task of eye blink identification in sign language data reasonably well, but only if supplemented by specific rules that take into account the temporal structure of eye blinks. However, we have also observed that there is a quite strong variation between individual videos/signers, so the solutions achieve very high results only when certain circumstances are favorable.

In most cases, the proposed CNN-based solution is performing slightly better than the EAR-based solution. Within the parameters of the CNN-based solution, using the eyes crops and training the CNN for 200 epochs, on data from 4 signers produces the best results. This can be taken into account in future studies.

Interestingly, of the two rules we proposed to account for the temporal structure of eye blinks, the simpler Rule 1 always performs best. It might be the case that the U-shape from Rule 2 is not an appropriate representation of the actual dynamics of eye lid movements, or that the CV/ML-based measurements are not precise enough to allow for this method to fully apply. Another explanation might lay in the chosen size of the window of frames which might not capture the full extent of a blink.

Note that both approaches of eye blink identification were tested with different threshold values for the CNN outputs or EAR, and the best results are reported. We also found that the optimal threshold values differ for the different videos and the different models. This can be explained for example by the fact that signers are holding their head down, therefore the threshold at which a blink may be observed is reduced, or by physiological differences between different people. This presents a complication for the practical use of these approaches for full automatic eye blink identification in novel data: for such an approach, specific threshold values must be provided to the model, and it might not be easy to determine in advance how to choose the value.

As discussed in Section 2.2, currently several other methods have been proposed for eye blink detection, but not specifically for sign language data, or with a linguistic definition of blinks in mind. We intend to test and adapt these approaches for further application to detecting blinks across sign languages.



## 6. Data availability

Codes and data are available on the OSF online repository: [https://osf.io/dth2q/?view\\_only=79f372c1776849258261c26f0f0b6ca6](https://osf.io/dth2q/?view_only=79f372c1776849258261c26f0f0b6ca6)

## Author Contributions

**Margaux Susman:** Conceptualization, Data Curation, Methodology, Formal Analysis, Investigation, Software, Writing. **Vadim Kimmelmann:** Conceptualization, Funding Acquisition, Writing

## Acknowledgements

Funded by the European Union (ERC, NONMANUAL, project number 101039378). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## 7. Bibliographical References

- C Baker and C Padden. 1978. Focusing on the nonmanual components of American Sign Language. In *Understanding language through sign language research*, p. siple edition, pages 27–57. Academic Press, New York.
- Valentin Belissen, Annelies Braffort, and Michèle Gouiffès. 2020. Dicta-Sign-LSF-v2: Remake of a Continuous French Sign Language Dialogue Corpus and a First Baseline for Automatic Sign Language Processing. In *12th conference on Language Resources and Evaluation*, Marseille, France.
- Anna Rita Bentivoglio, Susan B. Bressman, Emanuele Cassetta, Donatella Carretta, Pietro Tonali, and Alberto Albanese. 1997. *Analysis of blink rate patterns in normal subjects*. *Movement Disorders*, 12(6):1028–1034.
- Emilie Chételat-Pelé. 2010. *Les Gestes Non Manuels en Langue des Signes Françaises ; Annotation, analyse et formalisation : application aux mouvements des sourcils et aux clignements des yeux*. Theses, Université de Provence - Aix-Marseille.
- Emmanuel Descroix, Wojciech Świątkowski, and Christian Graff. 2022. *Blinking While Speaking and Talking, Hearing, and Listening: Communication or Individual Underlying Process?* *Journal of Nonverbal Behavior*, 46(1):19–44.
- Christine Dewi, Rung-Ching Chen, Xiaoyi Jiang, and Hui Yu. 2022. *Adjusting eye aspect ratio for strong eye blink detection based on facial landmarks*. *PeerJ Computer Science*, 8:e943.
- Ádám Fodor, Kristian Fenech, and András Lórinicz. 2023. *BlinkLinMulT: Transformer-Based Eye Blink Detection*. *Journal of Imaging*, 9(10):196.
- Grishchenko and V Bazaresvsky. 2020. *MediaPipe Holistic - Simultaneous Face, Hand and Pose Prediction, on Device*.
- A. Hall. 1945. *THE ORIGIN AND PURPOSES OF BLINKING*. *British Journal of Ophthalmology*, 29(9):445–467.
- Jeongmin Hong, Joseph Shin, Juhee Choi, and Minsam Ko. 2024. *Robust Eye Blink Detection Using Dual Embedding Video Vision Transformer*. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6274–6384.
- Paul Hömke, Judith Holler, and Stephen C. Levinson. 2017. *Eye Blinking as Addressee Feedback in Face-To-Face Conversation*. *Research on Language and Social Interaction*, 50(1):54–70.
- Bishar R. Ibrahim, Farhad M. Khalifa, Subhi R. M. Zeebaree, Nashwan A. Othman, Ahmed Alkhayyat, Rizgar R. Zebari, and Mohammed A. M. Sadeeq. 2021. *Embedded System for Eye Blink Detection Using Machine Learning Technique*. In *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*, pages 58–62, Babil, Iraq. IEEE.
- Craig N. Karson, Karen Faith Berman, Edward F. Donnelly, Wallace B. Mendelson, Joel E. Kleinman, and Richard Jed Wyatt. 1981. *Speaking, thinking, and blinking*. *Psychiatry Research*, 5(3):243–246.
- Oscar Koller. 2020. *Quantitative Survey of the State of the Art in Sign Language Recognition*. Publisher: arXiv Version Number: 2.
- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. *Gradient-based learning applied to document recognition*. *Proceedings of the IEEE*, 86(11):2278–2324.
- Silke Matthes, Thomas Hanke, Jakob Storz, Eleni Efthimiou, Athanasia-Lida Dimou, Panagiotis Karioris, Annelies Braffort, Annick Choisier, Julia Pelhate, and Eva Safar. 2010. *Elicitation tasks and materials designed for dicta-sign’s multi-lingual corpus*. In *sign-lang@ LREC 2010*, pages 158–163. European Language Resources Association (ELRA).

- Samuel G. Muller and Frank Hutter. 2021. [TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation](#). In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 754–762, Montreal, QC, Canada. IEEE.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [PyTorch: An Imperative Style, High-Performance Deep Learning Library](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Tran Thanh Phuong, Lam Thanh Hien, Do Nang Toan, and Ngo Duc Vinh. 2022. [An Eye Blink detection technique in video surveillance based on Eye Aspect Ratio](#). In *2022 24th International Conference on Advanced Communication Technology (ICACT)*, pages 534–538, PyeongChang Kwangwoon\_Do, Korea, Republic of. IEEE.
- Eric Ponder and W. P. Kennedy. 1927. [ON THE ACT OF BLINKING](#). *Quarterly Journal of Experimental Physiology*, 18(2):89–110.
- Chiarella Sforza, Mario Rango, Domenico Galante, Nereo Bresolin, and Virgilio F. Ferrario. 2008. [Spontaneous blinking in healthy persons: an optoelectronic study of eyelid motion](#). *Ophthalmic and Physiological Optics*, 28(4):345–353.
- Han Sloetjes and Peter Wittenburg. 2008. Annotation by category-elan and iso dcr. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.
- Terezq Soukupová and Jan Cech. 2016. [Real-Time Eye Blink Detection using Facial Landmarks](#). Rimske Toplice, Slovenia. Luka Cehovin, Rok Mandeljc, Vitomir Struc (eds.).
- John A Stern and June J Skelly. 1984. The eye blink and workload considerations. In *Proceedings of the human factors society annual meeting*, volume 28, pages 942–944. SAGE Publications Sage CA: Los Angeles, CA.
- F Sze. 2008. Blinks and intonational phrasing in hong kong sign language. In *Signs of the Time*, j. quer edition, pages 83–107. Signum, Hamburg.
- Ronnie Wilbur. 1994. [Eyeblinks & ASL Phrase Structure](#). *Sign Language Studies*, 84(1):221–240.
- Wenzheng Zeng, Yang Xiao, Sicheng Wei, Jinfang Gan, Xintao Zhang, Zhiguo Cao, Zhiwen Fang, and Joey Tianyi Zhou. 2023. [Real-time Multi-person Eyeblink Detection in the Wild for Untrimmed Video](#). ArXiv:2303.16053 [cs].

# SEDA: Simple and Effective Data Augmentation for Sign Language Understanding

Sihan Tan<sup>1,2</sup> , Taro Miyazaki<sup>2</sup> ,  
Katsutoshi Itoyama<sup>1,3</sup> , Kazuhiro Nakadai<sup>1</sup> 

<sup>1</sup>Tokyo Institute of Technology,

<sup>2</sup>NHK Science and Technology Research Laboratories,

<sup>3</sup>Honda Research Institute Japan Co., Ltd.

{tansihan, itoyama, nakadai}@ra.sc.e.titech.ac.jp, miyazaki.t-jw@nhk.or.jp

## Abstract

Sign language understanding (SLU) aims to convert sign language videos into glosses that transcribe sign language word-by-word by means of another written language and generate corresponding spoken sentences, including sign language recognition (SLR) and sign language translation (SLT). SLU has been a challenging undertaking since it demands the capability of fine-grained video understanding and sequence generation. In addition, the lack of supervised training data further hinders the advancement of SLU. To narrow the modality gap between vision and language and mitigate the data scarcity problem, we propose a **Simple and Effective Data Augmentation (SEDA)** framework for end-to-end SLU. In particular, SEDA consists of two key components: data augmentations on both sign and text sides and multi-task learning with task-specific fine-tuning. Experimental results on RWTH-PHOENIX Weather 2014T demonstrate that our proposed SEDA framework significantly and consistently outperforms the baseline model and achieves a WER of 19.91, a BLEU score of 25.19, and a ROUGE score of 51.72, delivering competitive scores in both SLR and SLT.

**Keywords:** Sign language understanding, Data augmentation, Multi-task learning.

## 1. Introduction

As the native language used by deaf and hard-of-hearing individuals to communicate, sign languages (SLs) exhibit distinctive grammar and have been established as a form of natural language (Klima and Bellugi, 1979). Sign language understanding (SLU) in which SLs are understood by means of machines mainly involves two functions: sign language recognition (SLR) and sign language translation (SLT). It is a challenging undertaking that requires the model to have the capability of fine-grained video understanding and sequence generation. Unlike spoken languages, SLs involve manual and non-manual elements (e.g., the movement of the body, head, mouth, or even eyebrows). Also, the visual signal in SLs displays dramatic variability among signers, posing a huge modality gap when transforming SLs into text (Zhang et al., 2023). Insufficient supervised training data presents an additional challenge to the advancement of SLU, as it increases the risk of overfitting. To tackle these challenges, it is essential to devise inductive biases, such as novel model architectures, training strategies and objectives, facilitating knowledge transfer, and the induction of universal representations for SLU. In this paper, we aim to augment SLs data on both sign and text sides, and provide effective training, including multi-task learning.

Existing SLU methods follow the framework of

neural machine translation (NMT) where the source language is spatial-temporal pixels rather than discrete tokens and the target language is spoken languages. Depending on the model architectures, annotation pairs, or final goals, SLU comprises: Sign2Gloss (Min et al., 2021; Hao et al., 2021), Sign2Gloss2Text (Yin and Read, 2020), Sign2(Gloss+Text) (Camgoz et al., 2020) and Sign2Text (Camgoz et al., 2018; Chen et al., 2022) tasks. Additionally, to boost the well-being of the sign language community and improve SLU performance, a number of studies have focused on Gloss2Text (Moryossef et al., 2021) and Text2Gloss (Miyazaki et al., 2020; Zhu et al., 2023) by transfer learning, data augmentation, etc. Following this line of study, we find that researchers seldom explore data augmentation techniques for the sign aspect, primarily concentrating on the textual component. Furthermore, constructing large-scale SL datasets with well-aligned annotations is a time-consuming and resource-consuming task (Uthus et al., 2023). For these reasons, developing a data augmentation technique for the sign side has become crucial.

In this paper, we propose a Simple and Effective Data Augmentation (SEDA) approach for SLU. The main idea is to increase the training samples and improve the model's performance by learning highly related tasks. Specifically, we adopt different sign embeddings to augment sign representations and

combine preprocessed spoken texts to achieve text augmentation. The contributions of this paper can be summarized as follows:

- We propose a Simple and Effective Data Augmentation (SEDA) approach to ease the data scarcity problem in the SLU task.
- Intensive analysis indicates that the SEDA method improves end-to-end SLU significantly through multi-task learning and task-specific fine-tuning.
- We evaluate the effectiveness of the proposed SEDA on the widely used dataset RWTH-PHOENIX Weather 2014T (PHOENIX14T). According to the experimental results, SEDA leads to notable enhancements in the SLU models, achieving competitive results in both SLR and SLT.

## 2. Related Work

### 2.1. Sign language understanding

In previous SLU research, the SLU methods can be divided into two categories: *cascading* and *end-to-end*. The cascading style adopts intermediate supervision, such as gloss, in which each gloss presents the manual transcription of a sign to convey its intended meaning. First, the sign language recognition model recognizes the continuous glosses from the unsegmented sign videos (**Sign2Gloss**), and then, the predicted glosses are utilized to generate the corresponding spoken sentence (**Gloss2Text**). In the Sign2Gloss system, a common architecture involves a feature extractor and a temporal modeling mechanism, such as Connectionist Temporal Classification (CTC) (Graves et al., 2006). However, most *cascading* SLU methods inevitably introduce an information bottleneck, as these methods utilize sign glosses as intermediate supervision. When the original sign video is transformed into glosses, some spatial-temporal information is lost (Yin and Read, 2020). End-to-end training is one promising way to achieve SLU. The *end-to-end* SLU directly converts the sign videos to spoken sentences (**Sign2Text**). Camgoz et al. (2018) formalizes this field by taking the SLU task as a neural NMT problem, demonstrating the practicality of the end-to-end method. In the following work (Camgoz et al., 2020), the sign glosses serve as the auxiliary supervision to regularize the neural encoder (**Sign2(Gloss+Text)**). Following this paradigm, we focus on the challenge of data scarcity by proposing a simple and effective data augmentation method. Besides, the data augmentation of sign language representation has rarely been explored before.

### 2.2. Multi-task learning

Multi-task learning aligns with the goal of increasing training samples and improving the model by learning related tasks (Zhang and Yang, 2018). Recently, the natural language processing (NLP) domain has benefited from adopting multi-task learning (e.g., multilingual translation). In Text2Gloss (Zhu et al., 2023), multilingual translation has been adopted to improve translation performance. As for SLU research, comprehensive multi-task learning experiments (i.e Sign2Gloss, Sign2Text, Gloss2Text, Text2Gloss, and MT) are conducted in (Zhang et al., 2023). These experiments offer valuable insights into how multi-task learning benefits SLT. We then combine multi-task learning with data augmentation and innovatively propose a simple and effective data augmentation (SEDA) framework for SLU. The following sections present the details of the proposed methods.

## 3. Methods

We applied the proposed SEDA to the sign language transformer (Camgoz et al., 2020) that widely serves as the baseline model in Sign2(Gloss+Text) and Sign2Text tasks. In this section, we first present the task definition and revisit the sign language transformer. We then give a comprehensive explanation of our proposed approaches, including data augmentation on both sign and text sides and multi-task learning.

### 3.1. Task definition

We formally define the setting of end-to-end Sign2(Gloss+Text). Given sign-gloss-text triples  $\mathcal{D} = (S_i, G_i, T_i)_{i=1}^N$ , where  $i$  and  $N$  represent the index of the input triple and the number of triples in the dataset,  $S_i = \{s_{i,z}\}_{z=1}^{|S_i|}$  denotes sign videos comprising  $|S_i|$  frames,  $G_i = \{g_{i,u}\}_{u=1}^{|G_i|}$  represents a gloss sequence with  $|G_i|$  gloss annotations, and  $T_i = \{t_{i,w}\}_{w=1}^{|T_i|}$  is the corresponding spoken text consisting of  $|T_i|$  words, and generally in SL data triples,  $|S_i| \gg |G_i|$  and  $|S_i| \gg |T_i|$ . The end-to-end Sign2(Gloss+Text) aims to predict glosses  $G$ , the text in sign order, and generate spoken text  $T$ .

### 3.2. Sign language transformers

The sign language transformer follows the encoder-decoder paradigm, with Transformer (Vaswani et al., 2017) as its backbone. It consists of five modules: a sign embedding, a transformer encoder, a CTC classifier, a word embedding, and a transformer decoder. In our sign language transformer, we introduce label smoothing to CTC training loss, aiming to mitigate the overfitting issue, and a new sign embedding to extract informative sign features.



**Sign embedding.** Replacing the sign embedding in (Camgoz et al., 2020), we adopt the re-trained one from self-mutual knowledge distillation (SMKD) model (Hao et al., 2021) followed by 1D-CNN layers to extract the informative sign representations. Here, we denote the new sign embedding as spatial-temporal embedding. During training, the parameters of the new pre-trained spatial-temporal embedding are frozen. We formulate this operation as:

$$f_i = \text{Spatial-temporalEmbedding}(S_i), \quad (1)$$

where  $f_i$  is the sign representation from the newly introduced spatial-temporal embedding.

**Transformer encoder.** The sign language transformer encoder intending to predict sign glosses  $G$  contains multi-layer transformer networks. The inputs of the transformer encoder are embedding sequences of tokens, such as the sign feature  $f_i$  from the spatial-temporal sign embedding. Unlike traditional seq2seq models, transformer networks do not employ recurrence or convolution mechanisms, which means they do not inherently contain positional information within sequences. To tackle this concern, we adopt the positional encoding introduced in (Vaswani et al., 2017) and add temporal order information into the embedded representations in the following manner:

$$\hat{f}_i = f_i + \text{PositionalEncoding}, \quad (2)$$

where the PositionalEncoding is pre-defined to generate a distinct vector in the shape of a sine wave that has been phase-shifted for each time step. Furthermore,  $\hat{f}_i$  is modeled using self-attention and projected into contextual representations  $h(S_i)$  that are fed forward to the transformer decoder to generate the target spoken text.

**CTC with label smoothing.** Sign language transformer employs glosses as auxiliary supervisions to train the transformer encoder. In the CTC-based Sign2Gloss tasks, CTC introduces the *blank* label, representing the silence or transition between two consecutive glosses. The extended glosses can be defined as  $\mathcal{G}^* = (g_{i,1}, \dots, g_{i,|G_i|}) \cup \{\text{blank}\} \in R^l$ , where  $l$  is the total number of labels. The CTC is utilized to compute the  $p(\mathcal{G}^*|h(S_i))$ , marginalizing over all possible  $h(S_i)$  to  $\mathcal{G}^*$  alignments as:

$$p(\mathcal{G}^*|h(S_i)) = \sum_{\pi \in \mathcal{B}} p(\pi|h(S_i)), \quad (3)$$

where  $\pi$  is a path and  $\mathcal{B}$  is the collection of all possible paths that lead to  $\mathcal{G}^*$ . The CTC loss in Sign2Gloss is defined as:

$$\mathcal{L}_{ctc} = 1 - p(\mathcal{G}^*|h(S_i)). \quad (4)$$

While CTC-based methods offer notable training convenience, as indicated in previous works (Min et al., 2021; Tan et al., 2023), they are prone to overfitting during training. Moreover, SLs are low-resource languages, this fact also poses the risk of overfitting. To mitigate the overfitting problem, we add a regularization term to the CTC objective function, which consists of the Kullback-Leibler (KL) divergence between the network’s predicted distribution  $P$  and a uniform distribution  $\mathcal{Q}$  over labels.

$$\mathcal{L}_{\mathcal{R}} = (1 - \alpha)\mathcal{L}_{ctc} + \alpha \sum_{t=1}^T D_{KL}(P||\mathcal{Q}) \quad (5)$$

Training the transformer encoder networks using CTC with label smoothing encourages the differences between the logits of the correct class and the logits of the incorrect classes to be a constant dependent on the weight ratio  $\alpha$ .

**Transformer decoder.** The sign language transformer decoder aims to generate the spoken sentence based on the contextual representation  $h(S_i)$ . It consists of cross-attention and self-attention layers. We introduce cross-entropy loss as the objective function of spoken language sentence generation.

$$\mathcal{L}_{\mathcal{T}} = - \sum_{u=1}^{|T_i|} \log \mathcal{P}(t_{i,u}|t_{i,<u}, h(S_i)) \quad (6)$$

### 3.3. Data augmentation

Data augmentation is a common technique used to relieve the risk of overfitting due to data scarcity. One commonly used data augmentation method involves original data with some minor changes. We apply the SEDA framework to the sign language transformer.

**Sign representation augmentation.** Instead of augmenting the sign frames directly, our proposed SEDA focuses on sign feature augmentation, that is, the same sign frames are processed by different sign embeddings. Given the sign video frames  $S_i = \{s_{i,z}\}_{z=1}^{|S_i|}$ , we propagate the  $S_i$  to different embedding layers (i.e., spatial embedding from (Camgoz et al., 2020) and newly introduced spatial-temporal sign embedding) separately to obtain multiple sign features. By taking this approach, we can obtain  $f_i \in \mathcal{F}$  from spatial-temporal embedding, where  $\mathcal{F} = \{f_1, f_2, \dots, f_N\}$ , and  $f'_i \in \mathcal{F}'$  from the original spatial embedding in the sign language transformer, where  $\mathcal{F}' = \{f'_1, f'_2, \dots, f'_N\}$ .

**Spoken text augmentation.** Inspired by the combining preprocessing methods in (Zhu et al., 2023),

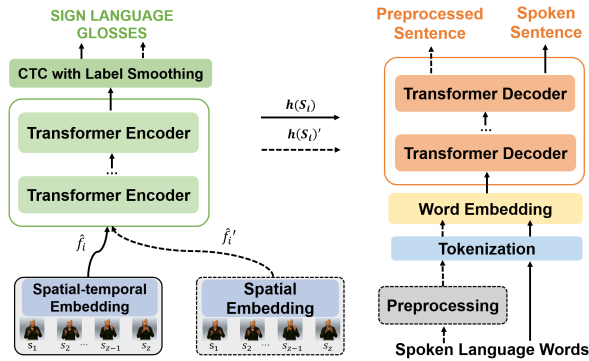


Figure 1: Overview of the proposed SEDA framework. The same sign frames will be forwarded to different sign embeddings to obtain multiple sign features. During multi-task learning, the sign features from spatial-temporal embedding are used to predict glosses and original spoken sentences, as shown by the solid line. Meanwhile the sign features from spatial embedding are fed to the model to generate glosses and preprocessed sentences, which is presented by the dotted line.

we apply preprocessing techniques to the spoken sentence  $T_i$ . We conduct lemmatization and alphabet normalization on the PHOENIX14T dataset (Camgoz et al., 2018) and combine the processed data with the original annotations. Lemmatization is the linguistic process of reducing words to their base or root form. Alphabet normalization is employed to convert specific letters, such as ü, ö, ä, and ß, into their corresponding counterparts. The processed spoken text, denoted as  $T'_i$ , is then paired with the copied gloss sequence  $G_i$  to become a new training dataset on the target side. Once the augmented sign features and preprocessed spoken text annotations are obtained, we are able to construct new data triples  $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^N$  and  $\mathcal{D}_2 = (f'_i, G_i, T'_i)_{i=1}^N$ .

### 3.4. Multi-task learning

The augmented data triples, represented as  $\mathcal{D}_1$  and  $\mathcal{D}_2$ , are then mixed up and fed to the sign language transformer one after another. As shown in Fig. 1, when presented with the input  $f_i$ , the sign language transformer encoder is trained to predict  $G_i$ , and the transformer decoder is trained to generate  $T_i$ . The same procedure applies to the input  $f'_i$ , where the sign language transformer encoder predicts  $G_i$ , and the transformer decoder generates  $T'_i$ .

The networks are trained by minimizing the joint loss term  $\mathcal{L}$ , which is the weighted sum of the translation loss  $\mathcal{L}_T$  and the gloss prediction loss  $\mathcal{L}_R$ , as follows:

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_T \mathcal{L}_T, \quad (7)$$

where  $\lambda_R$  and  $\lambda_T$  are hyperparameters that de-

termine recognition and translation loss function weight during training. Since our final goal is to predict  $G_i$  and generate  $T_i$ , we then fine-tune the network using  $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^N$ .

## 4. Experiments

### 4.1. Experimental setup

To evaluate the effectiveness of the proposed SEDA framework, we conducted ablation experiments.

**Model setting.** For training hyper-parameters, we start mainly from the setting for the sign language transformer<sup>1</sup>. In particular, we keep  $\alpha = 0.01$ ,  $\lambda_R = 5.0$ , and  $\lambda_T = 1.0$ , which is empirically decided. As suggested in (Zhu et al., 2023), the model performance increases when the number of encoders or decoders is reduced compared to the original transformer architecture in SL translation scenarios. We performed extensive experiments. As the results indicated, we maintained the encoder depth at 2 and the decoder depth at 4.

**Dataset.** We worked on the widely utilized PHOENIX14T dataset and augmented the spoken texts (Zhu et al., 2023). The details of the augmented information are shown in Table 2. Note that we used  $\mathcal{D}_1$  for the development and test.

**Evaluation metrics.** We report the experimental results mainly on the Sign2 (Gloss+Text) task, including the Sign2Gloss and the total Sign2 (Gloss+Text) results. The most common measure of Sign2Gloss performance is the word error rate (WER), which can be calculated as:

$$\text{WER} = \frac{S + D + I}{S + D + C}, \quad (8)$$

where  $S$ ,  $D$ ,  $I$ , and  $C$  indicate the number of **S**ubstitutions, **D**eletions, **I**nsertions, and **C**orrections, respectively. For SLT task, we use standard metrics commonly used in machine translation, including tokenized BLEU (Papineni et al., 2002) with 4-grams and the Rouge-L F1 (ROUGE) (Lin, 2004).

### 4.2. Experimental results

We evaluated the proposed SEDA framework on augmented PHOENIX14T. The main results are listed in Table 1. On the PHOENIX14T development set, the proposed SEDA surpassed the baseline by 9.93 WER, 4.24 BLEU, and 4.65 ROUGE. It also outperformed the state-of-the-art end-to-end or cascading approaches.

### 4.3. Discussion

**Introducing high-quality spatial-temporal sign embedding improves SLR and SLT.** Replacing

<sup>1</sup><https://github.com/neccam/slt>

Table 1: End-to-end SLU performance on PHOENIX14T dataset

Methods	DEV			TEST		
	WER↓	BLEU-4↑	ROUGE↑	WER↓	BLEU-4↑	ROUGE↑
<b>Previous research (end-to-end)</b>						
Joint-SLRT (Camgoz et al., 2020)	24.98	22.38	–	26.16	21.32	–
Sign Back Translation (Zhou et al., 2021)	22.70	23.90	50.29	24.45	24.34	49.54
STMC-T <sup>*</sup> (Zhou et al., 2022)	–	24.09	48.24	–	23.65	46.65
<b>Previous research (cascading)</b>						
STMC-Transformer <sup>*</sup> (Yin and Read, 2020)	<b>19.60</b>	22.47	48.70	<b>21.00</b>	22.47	48.78
<b>Ours (end-to-end)</b>						
Baseline	29.84	20.95	47.07	28.67	21.70	47.82
+ High-quality Spatial-temporal Sign Embedding	21.40	22.28	48.81	22.59	22.86	48.97
+ CTC Label Smoothing	21.56	23.05	48.86	22.05	22.40	47.58
+ Multi-task Learning	20.36	23.88	50.57	21.79	23.34	49.71
+ Fine-tune	<b>19.91</b>	25.19	<b>51.72</b>	<b>21.51</b>	<b>24.89</b>	<b>51.61</b>
+ Gloss-less fine-tune	–	<b>25.35</b>	51.40	–	24.75	50.77

\* denotes using extra clues (keypoints)

Table 2: Statistics of preprocessed PHOENIX14T

	Original Text			Preprocessed text		
	Train	Dev	Test	Train	Dev	Test
Instance	7,096	519	642	7,096	519	642
Vocab.	1,066	393	411	2,216	793	836
tot. words	99,081	6,820	7,816	99,081	6,820	7,816
tot.OOVs	–	57	60	–	39	38

the original sign embedding, we introduce the re-trained one from the SKMD model and adopt 1D-CNN layers to extract the spatial-temporal sign information. This replacement delivers notable enhancements in both SLR and SLT (−8.44 WER, +1.33 BLEU, +1.74 ROUGE on the dev set). Adding a regularization term to the CTC, we observe an improvement in SLT (+ 0.77 BLEU on the dev set).

### Multi-task learning enhances both SLR and SLT.

The sharing of parameters through multi-task learning, using augmented dataset, facilitates knowledge transfer. As shown in Table 1, the multi-task learning achieves a quality boost (−1.2 WER, + 0.83 BLEU, +1.71 ROUGE on the dev set). Sharing the mixed parameters benefits tasks but lacks of task-specific characteristics. For this, we performed fine-tuning in the following.

**Mixing shared parameters with task-specific parameters further provides quality gains.** We further conduct task-specific fine-tuning using the data triples  $\mathcal{D}_1 = (f_i, G_i, T_i)_{i=1}^N$ . Here gloss-less fine-tuning refers to using the multi-task learning applied model, and we fine-tune the model to do the Sign2Text task without glosses. By task-specific fine-tuning, SLR and SLT tasks undergo a dramatic improvement (Fine-tune: −0.45 WER, + 1.31 BLEU, + 1.15 ROUGE; Gloss-less fine-tune: +1.47 BLEU, +0.83 ROUGE on the dev set).

## 5. Conclusion

In this paper, we propose a simple and effective data augmentation (SEDA) method to mitigate the data scarcity problems in end-to-end sign language understanding (SLU). The SEDA approach includes adopting different sign embeddings, combining preprocessed spoken texts, and a multi-task learning strategy. The former two methods increase the amount of training data, especially the sign representations, which has rarely been conducted before. Multi-task learning narrows the gap between vision and language by sharing mixed parameters. Experimental results on the widely utilized PHOENIX14T dataset indicate that our proposed SEDA benefits the end-to-end SLU, surpassing the baseline by 9.93 WER, 4.24 BLEU score, and 4.65 ROUGE score and achieving competitive results in both sign language recognition (SLR) and translation (SLT) tasks.

## 6. Limitations

While our SEDA framework significantly benefits the end-to-end SLU on the PHOENIX14T dataset, it still faces the limitation that more datasets, such as the German sign language dataset (Public DGS Corpus (Hanke et al., 2020)) or Chinese sign language dataset (CSL-Daily (Zhou et al., 2021)), are needed to demonstrate the universality of the proposed method. We will adopt multiple datasets and conduct more detailed analyses in future work.

## 7. Acknowledgements

This work was supported by JSPS KAKENHI Grant No. JP19KK0260, JP20H00475 and JP23K11160.

## 8. Bibliographical References

- Necati Cihan Camgoz et al. 2018. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Necati Cihan Camgoz et al. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yutong Chen et al. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5120–5130.
- Alex Graves et al. 2006. [Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks](#). In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, page 369–376, New York, NY, USA. Association for Computing Machinery.
- Thomas Hanke et al. 2020. [Extending the Public DGS Corpus in size and depth](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 75–82, Marseille, France. European Language Resources Association (ELRA).
- Aiming Hao et al. 2021. Self-mutual distillation learning for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11303–11312.
- Edward S Klima and Ursula Bellugi. 1979. *The signs of language*. Harvard University Press.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yuecong Min et al. 2021. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11542–11551.
- Taro Miyazaki et al. 2020. [Machine translation from spoken language to sign language using pre-trained language model as encoder](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).
- Amit Moryossef et al. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.
- Kishore Papineni et al. 2002. [Bleu: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.
- Sihan Tan et al. 2023. Improving sign language understanding introducing label smoothing. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 113–118. IEEE.
- David Uthus et al. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. *arXiv preprint arXiv:2306.15162*.
- Ashish Vaswani et al. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.
- Biao Zhang et al. 2023. Sltunet: A simple unified model for sign language translation. *arXiv preprint arXiv:2305.01778*.
- Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review*, 5(1):30–43.
- Hao Zhou et al. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1325.
- Hao Zhou et al. 2022. [Spatial-temporal multi-cue network for sign language recognition and translation](#). *IEEE Transactions on Multimedia*, 24:768–779.
- Dele Zhu et al. 2023. [Neural machine translation methods for translating text to sign language glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.



# HamNoSys-based Motion Editing Method for Sign Language

Tsubasa Uchida , Taro Miyazaki , Hiroyuki Kaneko 

NHK Science & Technology Research Laboratories  
Tokyo Japan  
{uchida.t-fi, miyazaki.t-jw, kaneko.h-dk}@nhk.or.jp

## Abstract

We have developed a Japanese-to-Japanese Sign Language (JSL) translation system to expand sign language services for the Deaf. Although recording the motion data of isolated JSL by motion capture (MoCap) and avatar animation driven by MoCap data is effective for capturing the more natural movements of sign language, the disadvantage is that they lack the flexibility to reproduce the contextual modification of signs. We therefore propose a sign language motion data editing method based on the Hamburg Notation System for Sign Languages (HamNoSys) for use in a hybrid system that combines a MoCap data-driven technique and a phonological generation technique. The proposed method enables the editing of handshape, hand orientation, and location of the motion data based on HamNoSys components to generate contextual modifications for motion-captured citation form signs in translated gloss sequences. Experimental results demonstrate that our method achieves the flexibility to generate contextual modifications and new movements while preserving natural human-like movements without the need for additional MoCap processes.

**Keywords:** HamNoSys, Avatar Animation, Motion Capture, Japanese Sign Language (JSL), Classifier

## 1. Introduction

In order to improve accessibility for the Deaf, the number of services with sign language (SL) has been gradually increasing in recent years. When it comes to expanding SL services, translation from spoken language to SL is the most basic and important requirement for Deaf people. The best way to provide information in SL is through real-time interpretation by a human signer, and in some cases using pre-recorded video of signers. The pre-recorded video method allows for higher quality SL interpretation than real-time because it can be prepared in advance, but creating a signed video from spoken text or audio requires recording time and cost. Additionally, as human video material of interpreters signing is difficult and inflexible to reuse, signers are required to translate and record new videos each time they create new content. There is also another issue regarding the anonymization of signers in video material (Xia et al., 2022).

Therefore, several automatic translation methods from spoken language to SL have been proposed to increase the number of signed videos without human intervention. Recently, with the development of deep neural network technology, end-to-end translation methods that generate photo-realistic SL videos from input spoken language have been proposed (Saunders et al., 2020, 2022). However, even though hands and body movements are reproduced in the output video, facial expressions and mouth shape cannot be completely reproduced. At present, the mainstream output of SL translation comprises avatar animation created

using computer graphics (CG) technology. There are several methods when it comes to producing an avatar animation as an output of SL translation, such as hand-crafted keyframe animation (McDonald et al., 2016), phonological-based generation (Nunnari et al., 2018; McDonald and Filhol, 2021), and data-driven methods using motion capture (MoCap) data (Gibet et al., 2011; Naert et al., 2020; Brock and Nakadai, 2018). Since the motion data-driven method can reproduce natural movements that are more realistic than other methods, we have developed a Japanese Sign Language (JSL) translation system that utilizes pre-recorded MoCap data (Miyazaki et al., 2023; Uchida et al., 2023).

SL translation systems mainly use utterance synthesis to concatenate sign words together for accurately representing sentences in spoken languages (Kim et al., 2022; Ebling, 2016; Morrissey, 2008). In utterance synthesis, it is not enough to simply concatenate an SL word in a citation form because in SL, word forms such as handshape, location, speed, and size change depending on the context. Therefore, to generate accurate SL sentences, it is essential to reproduce the modification of signs according to the context of the original spoken sentences (Naert et al., 2020). We are currently developing an editing tool to reproduce contextual modifications of JSL by modifying MoCap data captured in a citation form (Uchida et al., 2023). The term “contextual modification” is used in this paper to refer to modifications of signs based on the linguistic context in the broad sense, such as prosody, assimilation, and morphological modifications.

In this paper, we present a motion editing method

based on the Hamburg Notation System for Sign Languages (HamNoSys) (Hanke, 2004). Our contributions are summarized as follows:

1. We developed a motion editing method specialized for MoCap data-driven SL avatar animation.
2. Our method can edit the handshape, hand orientation, and location of pre-recorded MoCap data based on the linguistic components of HamNoSys.
3. Evaluation experiments showed that our method can output natural human-like movement and offers the flexibility to generate contextual modifications and new movement to improve the intelligibility of avatar animations without additional MoCap processes.

## 2. Related Work

Approaches to SL avatar animation production include hand-crafted keyframe animation, phonological-based generation, and data-driven methods. Each method has its own advantages and disadvantages, and the perfect generation method has yet to be established (Wolfe et al., 2022). All three methods mainly utilize utterance synthesis, which generates SL sentences by concatenating isolated citation form signs.

The first hand-crafted keyframe animation method generates avatar animation by interpolating the motion between the key poses of the skeleton (McDonald et al., 2016). This method has the advantage of being able to generate motions for SL sentences using a small amount of pose data, and it can flexibly respond to natural contextual modifications by creating poses with various patterns. The disadvantage of this approach is that the quality and quantity of the output animation depends on the quality and the quantity of manual labor. All poses have to be created by hand by animators in advance, requiring a large amount of work from highly skilled animators to achieve high quality and varied translations.

The second phonological-based generation method can generate motions that reproduce contextual modifications in SL utterances by taking linguistic information into account. Some researchers are developing a system that parametrically generates avatar motion by combining phonological components based on SL description methods proposed through SL linguistic research, such as Stokoe et al. (1965); Hanke (2004).

This approach allows the system to freely generate motions by selecting and combining multiple phonological parameters, so there is no need to prepare motion data in advance. It is also extremely

flexible for reproducing all contextual modifications in an utterance (Elliott et al., 2007; Fotinea et al., 2007; McDonald and Filhol, 2021). However, there is a problem in that the generated motions are typically robotic and unnatural. Furthermore, in order to use this animation generation approach in an SL translation system, there is another problem in that very detailed annotations by linguistic experts who are familiar with the structure of each description method are required when preparing training data in advance.

The final method utilizes MoCap data to manipulate avatars for generating realistic human movements. Since this method captures human movements as motion data, it can reproduce the most natural human-like SL movements among the three methods discussed here. However, MoCap recording requires large-scale studio equipment and post-processing of recorded data by experts, making this approach expensive and time-consuming. Additionally, it is difficult to reproduce contextual modifications using only motion data recorded in citation form in utterance synthesis, and it is also impossible to record the almost infinite number of SL modifications in all the possible contexts as motion data.

Therefore, some researchers have proposed hybrid methods that combine the advantages of these methods.

Huenerfauth et al. (2015) proposed modeling methods for the construction of a parameterized lexicon of ASL verb signs. They modeled the signer's hand locations and orientations during each ASL indicating verb, dependent upon the location in the signing space where the subject and object were positioned, and utilized SL MoCap data from native signers as a training data set for learning the models.

Filhol and McDonald (2018) proposed a hybrid system of the hand-crafted keyframe animation system Paula and SL description model AZee. The system can produce avatar animations that can be modified procedurally thanks to Inverse Kinematics (IK) solvers in order to synthesize proforms or spatial referencing mechanisms for utterance generation. Furthermore, Nunnari et al. (2018) proposed a bottom-up procedural computation method for the AZee animation system, allowing for the animation of different communicative channels that are interleaved on the timeline.

Gibet et al. (2011) and Naert et al. (2021) proposed a hybrid method that combines the naturalness of MoCap data-driven methods with the flexibility of the phonological-based generation method. In this method, MoCap data divided into several element channels, (e.g., handshape, position, movement, posture, and facial expression) is registered as a MoCap data corpus in advance, and the data

is combined to create new SL utterances. However, this approach requires building a semantic database that serves as a mapping between the SL gloss level annotations and the movements in the MoCap data corpus, and it requires a very extensive pre-process (such as detailed segmentation) on the recorded MoCap data.

Although these methods have potential as elemental technologies for SL avatar animation production, few efforts have been made to utilize them as part of a translation system. Therefore, we propose a system that combines a MoCap data-driven method and a flexible phonological-based generation method as a motion editing tool for translation systems. The proposed system reproduces natural contextual modifications for the avatar animation that is the output of the translation system by editing the SL movement of MoCap data using HamNoSys-based phonetic units. Furthermore, it is also possible to generate new SL motions from existing MoCap data without any high-cost MoCap processes.

### 3. Data Preparation

#### 3.1. Motion capture

We captured over 8,000 isolated JSL motion data by MoCap. Among the various motion capture methods that have been proposed, we opted to use an optical motion capture system, which is expected to have high accuracy. In this system, multiple markers coated with retroreflective material are attached to the signer, and reflected light is recorded utilizing multiple infrared cameras equipped with infrared LEDs around the lenses. The motion of the signer is calculated by measuring the three-dimensional position of each marker based on the recorded video. We used 42 Vicon T160 and V16 cameras and 112 retro-reflective markers on the signer’s body (Table 1), which have less physical restraint than sensor gloves (Figure 1). We used significantly more cameras and markers than in general movie production shooting, because in SL, details such as whether the ring formed by the fingers is closed or not, and whether adjacent fingers are touching each other are important for understanding. Of course, in order to record non-manual markers (NMMs) such as facial expressions, which are important in JSL as well, we also recorded motion data by attaching markers to characteristic parts of the signer’s face. For the handshape and facial expression, we used markers sized 3 mm in diameter because they need to be measured with high precision. Most of this data was captured as citation forms listed in the JSL dictionary, and any data loss due to marker occlusion, which is a weak point of the optical motion capture system, was

carefully corrected during post-processing. The MoCap data was recorded in Filmbox (FBX) format at 120fps and post-processed into Biovision Hierarchy (BVH) format files at 60fps. BVH is a MoCap file format developed by Biovision, and is written in ASCII format, making it easy to manipulate the file contents on a computer.

In addition, we constructed a JSL motion database that can be utilized as an independent API. The database stores motion data and JSL gloss pair information. The translation system reads each linked motion data from the motion database based on the JSL gloss string that is the translation result. The motion database also provides start and end frame information for each motion data, so unnecessary frames can be cut when using utterance synthesis.

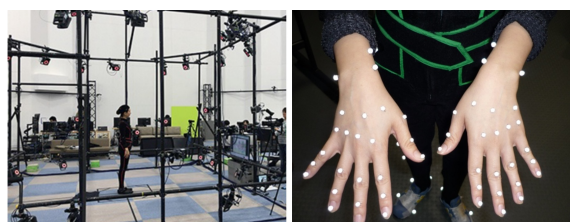


Figure 1: Optical motion capture system and markers for the JSL recording.

Body region	Retro-reflective markers	
	Diameter [mm]	Number
Face	3	33
Hands	3	24 × 2
Body	10	31
Total number		112

Table 1: Our motion capture marker set.

#### 3.2. Bone structure of avatar

We created a CG avatar with a skeletal structure that can appropriately reproduce the obtained MoCap data (Kaneko et al., 2010). Our avatar has a total of 162 joints (one root joint, 111 body and facial joints, and 50 end-effectors). In defining the skeletal structure, we made it possible for the fingers to make fine movements at the base of the thumb and palm in addition to the joints of each finger, and for the body, we added joints to the collarbone to enable movement around the shoulders (Figure 2).

One technique for expressing facial expressions using CG is to define the facial muscles under the skin of a CG avatar’s face and the facial skin surface that are linked to these as a physical model, but creating a facial muscle model is extremely difficult

and complex. As an alternative, we implemented a method that controls the skin shape of a CG avatar using data that digitally records the movement of the facial skin shape using optical motion capture data. Even though it has more controllable points (joints) than the facial muscle model, it does not require simulation using a physical model, it reduces the computational load during CG animation generation, and it is easy to create a CG avatar's face.



Figure 2: Our avatar representation driven by MoCap data.

## 4. Proposed Method

### 4.1. Overview

Figure 3 shows an overview of the avatar-based Japanese-to-JSL translation system we developed. Our system translates Japanese text into a JSL gloss sequence by means of a transformer model (Miyazaki et al., 2023), then performs utterance synthesis by concatenating JSL motion data, and finally displays it as a JSL avatar animation video on a player developed on the Unity game engine platform (Unity Technologies, Accessed: 2024-02-29). We also developed a motion editing tool on the Unity platform as a function linked to the motion blending process shown in Figure 3 (Uchida et al., 2023). This tool can fix translation errors by manually replacing motion data, adjusting the motion speed and connection interval, etc. before performing utterance synthesis. We added HamNoSys-based editing, which is the method proposed in this paper, to reproduce contextual modifications from citation form MoCap data as one of the new functions of this motion editing tool.

HamNoSys is the notation system developed by the University of Hamburg as a means of transcribing signs on a phonetic level. It transcribes manual postures and movements in signs based on four components (handshape, hand orientation, location, and action), and the latest version, HamNoSys 4.0, also takes into account NMMs such as eye gaze, facial expressions, and mouth gestures

(Hanke, 2004). It is a sufficiently general model of sign language phonetics that all sign languages can be transcribed. Therefore, it is also possible to create motions for other sign languages by using our proposed method and JSL motion data.

As a first stage of implementation, we have adopted three components of HamNoSys, namely, handshape, hand orientation, and location, for the motion editing function. Since there are many variations of SL actions, we implemented our proposed method by specifying the motion data that is the source of editing, using that movement as an action, and applying the other three components. The proposed method can be used not only for contextual modification editing in a JSL translation system but also as an individual tool for generating new motion data without any MoCap process. We also developed a user-friendly GUI for non-experts of HamNoSys transcription rules, where operators can intuitively select the handshape, hand orientation, and location by referring to illustrations.

Each of the three functions (handshape, hand orientation, and location) are explained below along with editing examples.

### 4.2. Handshape

The handshape function replaces the avatar's handshape in the motion being edited with handshape motion data prepared in advance. All rotation information of the joints from the wrist of the avatar onwards in the motion data to be edited is replaced with the rotation information of the corresponding joints of the prepared handshape motion. We prepared several types of handshape motions based on HamNoSys's handshape chart. The basis posture data for the handshape was created by selecting a motion that included the relevant handshape from the existing JSL motion database, and cutting out the keyframes. Figure 4 shows the handshape list GUI and an example of generating a JSL motion [TOKYO TOWER] by replacing only the handshape from a JSL motion [SKYTREE] using the proposed method. The operator can select which handshape to replace from the handshape illustration list.

### 4.3. Hand orientation

The hand orientation function rotates the avatar's wrist to change the orientation of the palm. This function rotates the palm by changing the value of the avatar's wrist joint rotation data according to the rotation angle in eight patterns defined by HamNoSys. Also, the function can change which direction to use as the rotation axis from the 18 defined directions when rotating the palm. Figure 5 shows the hand orientation GUI and an example of generating a JSL motion [EMAIL] by changing only hand orientation from a JSL motion [PAY] using the



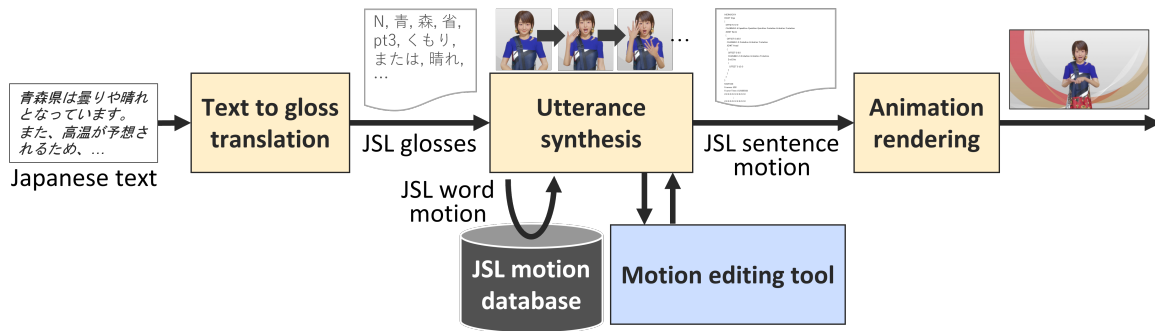


Figure 3: Japanese-to-JSL translation system.

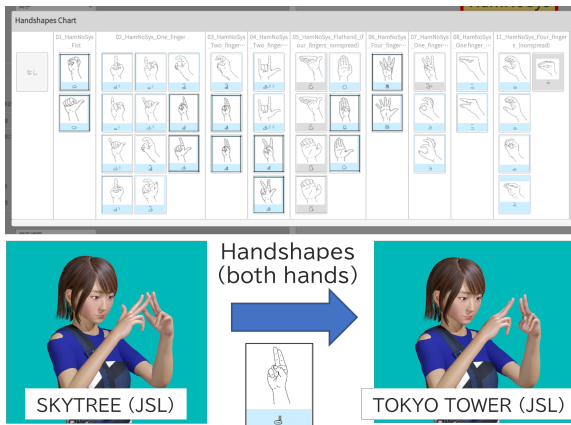


Figure 4: Handshape replace function.

proposed method. The operator can select which combination of direction and orientation to change from the hand direction and orientation illustration.

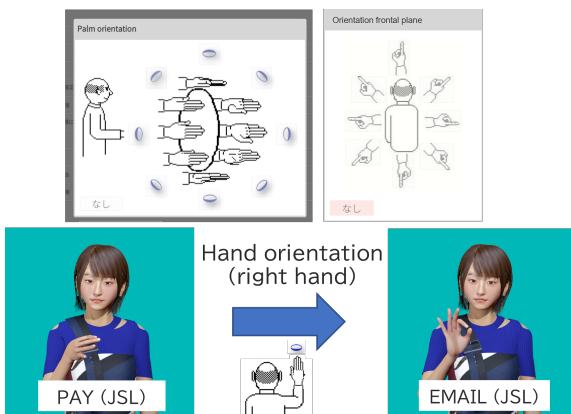


Figure 5: Hand orientation control function.

#### 4.4. Location

The location function changes the location of the expression of signs. This function moves the location of signs to a specific location defined by HamNoSys by changing the rotation information of the avatar's

arms using IK.

We used an IK articulated chain for the sign's location change. The chain consists of four joints (shoulder, elbow, forearm, and wrist) for each arm. We initially used Cyclic Coordinate Descent Inverse Kinematics (CCDIK) as an IK solver, but the hands could not reach the target position after the change process. Naert et al. (2021) proposed using Forward And Backward Reaching Inverse Kinematics (FABRIK) (Aristidou and Lasenby, 2011) for modification of hand placement. In contrast to CCDIK, which is a method that fixes the root joint position and iteratively updates the joint rotation, FABRIK is a heuristic method that obtains an IK solution by repeatedly adjusting joint angles while alternately using the root and end-effector as reference points. Figure 6 shows the difference in location change between the two types of IK solvers. By replacing CCDIK with FABRIK, the avatar's arms could be extended and reach closer to the target after the location change. Therefore, we adopted FABRIK as the new IK solver. Of course, each IK algorithm has its advantages and disadvantages, so we believe that continued consideration is necessary.

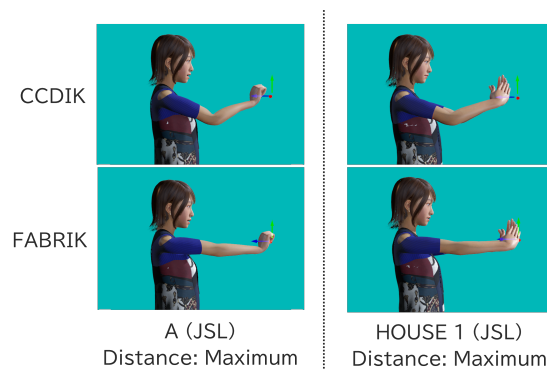


Figure 6: Comparison of IK solvers: CCDIK (upper) and FABRIK (lower).

Furthermore, depending on the original SL motion to be edited, a problem arises in that the linguistically meaningful hand configuration before editing

collapses after the location is changed. This is because the rotation of the wrist is linked to changes in the posture of the arms.

Therefore, we developed a new method to reproduce the original hand configuration represented in citation form after the location change. This method offsets the rotation value of the wrists by rotating the wrists in the opposite direction after movement in accordance with the amount of rotation of the joints of the arms, thereby reproducing the configuration of the signs before editing. Figure 7 shows before and after images of applying the wrist offset method.

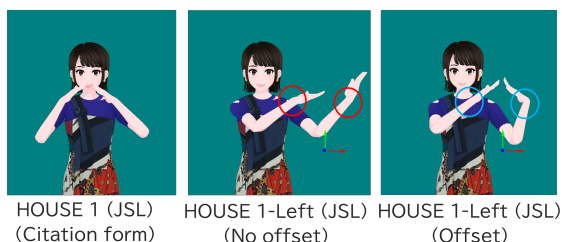


Figure 7: Before and after applying wrist offset method. Left: Citation form. Middle: No offset. Right: Offset.

By incorporating FABRIK and the wrist offset method, the range of contextual modifications of signs that can be reproduced by the location change function has been expanded. Figure 8 shows the location GUI and an example of contextual modification of a JSL motion [HOUSE 1] by changing its location using the proposed method. The operator can change the location of signs by choosing the illustration of the location relative to the avatar's body and face.

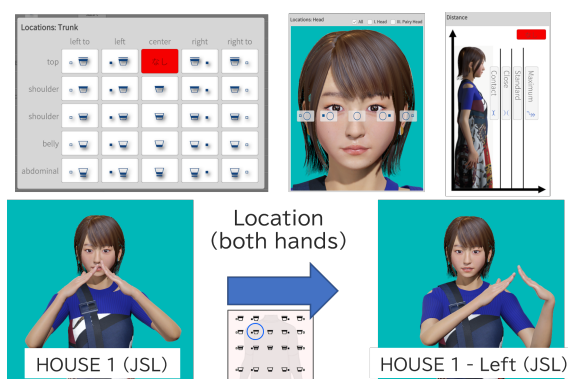


Figure 8: Location change function.

#### 4.5. Example of motion editing by proposed system

An example of contextual motion modifications is shown in Figure 9. The GUI of the motion edit-

ing tool has a function to visually connect isolated motion data as nodes and edit the parameters of each motion data, and we also added a node dedicated to HamNoSys editing. By connecting the HamNoSys editing node to the motion data to be edited, each of the three components—handshape, hand orientation, and location—can be edited independently.

Using the proposed method, we replaced the right handshape of JSL [GO 4] and rotated the wrist in the translated JSL gloss sequence to produce a more natural JSL animation that clearly expresses the means of going and the number of people.

This usage is linguistically called a classifier (CL) predicate, and is one of the contextual modifications that can be reproduced using the proposed method. Some researchers have also worked on reproducing this CL predicate in avatar animation (Huenerfauth, 2006; Filhol and McDonald, 2020; Naert et al., 2021). Since our method is based on MoCap data, it is possible to reproduce more realistic CL predicate motions in JSL translation results than a method that uses only a phonological-based generation technique.

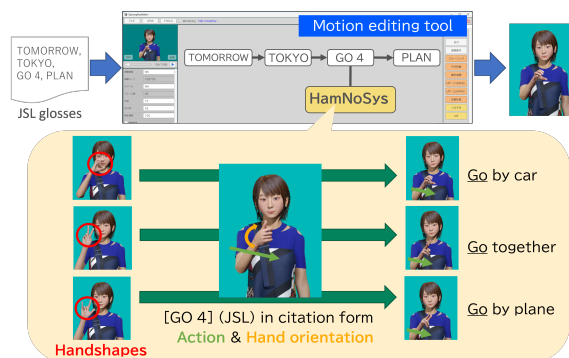


Figure 9: Example of contextual motion modifications for CL predicate.

An example of generating a German Sign Language (DGS) motion from JSL motion using our proposed method is shown in Figure 10. The upper part of the figure is an example of generating DGS motion [WICHTIG 1] by replacing only handshape from a JSL motion [STUDY 2], and the lower part is an example of generating DGS motion [SAGEN 1] by changing only hand orientation from a JSL motion [SAY 1]. As demonstrated in this example, it is also possible to generate new motions in other sign languages from JSL motions.

## 5. Evaluation

### 5.1. Design of the evaluation

We conducted an evaluation experiment on JSL avatar animations generated by our JSL translation

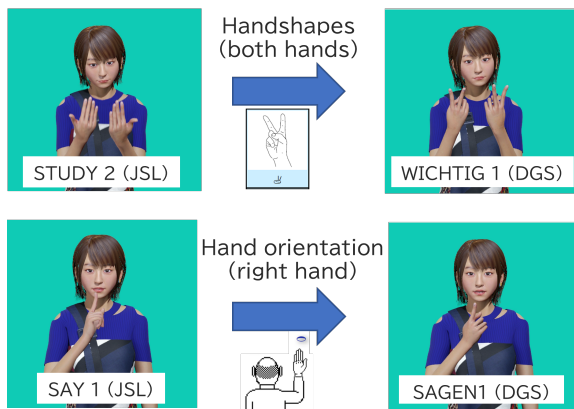


Figure 10: Example of DGS motion generation from JSL motion.

system implemented using the proposed editing method based on HamNoSys. We recruited four participants, three men and one woman, for this experiment. Two of the participants were born Deaf, one was hard of hearing, and one was a child of Deaf adults (CODA).

To investigate the effect of modifying that combines two functions, namely, handshape and hand orientation, we prepared and compared JSL avatar animations with and without modifying for JSL sentences containing CL predicates. To generate the videos used for evaluation, we selected ten Japanese news sentences that include CL predicate expressions from our Japanese-JSL news corpus. We prepared a total of 20 avatar animation videos: ten that were automatically generated by inputting the ten selected Japanese news sentences into our translation system, and ten that were manually modified using our proposed method after being automatically generated. The modification was carried out by replacing only the handshape and changing the hand orientation for the citation form motion of the word corresponding to the CL predicate. An example of modification is to change the handshape and hand orientation for the motion data of CL predicates such as [GO 4] (as shown in Figure 9), [MEET 7], [HELPED 1], and [PROTECT 1]. By modifying the JSL expressions to match the means and number of people in the original Japanese context, we aimed to clarify the subject in the JSL sentences and improve understanding of the content.

All participants evaluated all 20 avatar animation videos in the experiment. The number of video views was unlimited. The videos were presented in random order, regardless of whether they were modified by the proposed method.

Participants answered three questions after watching each video: a question testing the intelligibility of the JSL sentence, a question on the accuracy of the JSL expression, and a question about

the realism of the utterance synthesis produced. All questions and answers were conducted through JSL by a JSL interpreter for Deaf and hard of hearing persons, and directly in spoken Japanese for the CODA person.

First, to check the intelligibility of the JSL sentences, we asked the subjects: “Please tell us what you were able to know by watching the animation.” This was done to determine whether they understood the JSL expressions related to CL predicates correctly in context. The second and third questions were based on questions used in the evaluation of previous studies (Naert, 2020). The second question concerned the accuracy of the JSL: “Do you think that the sign was done correctly?”, and the third question evaluated the naturalness of the movement: “Do you think that the sentence in JSL is natural/realistic/spontaneous (does it seem like the movement of a real person)?”. Both questions were answered on a 5-point Likert scale ranging from 1 (most negative) to 5 (most positive), as in previous studies.

## 5.2. Results

We defined the recognition rate as the percentage of people who correctly understood the meaning of the CL predicate part in each JSL sentence according to the context, and is the average value of the four participants. Figure 11 shows the recognition rate of the CL predicate part for each JSL sentence. Out of a total of ten sentences, the recognition rate of four sentences was improved after modifying by the proposed method (SL3, SL4, SL6, SL9), 3 sentences remained unchanged (SL1, SL5, SL10), and the remaining three sentences could be recognized correctly with or without modifying (SL2, SL7, SL8).

For example, in SL3, the handshape with the thumb up in the JSL for [GO 4] automatically generated by the translation system was manually replaced with a handshape with two fingers raised, making it more clear that two people are going. Similarly, in SL4, the handshape with the index finger raised in the JSL for [GO 2] was replaced with a handshape with three fingers raised, making it more clear that three people are going. Also, in SL9, the handshape representing the CL of the airplane in the JSL for [LANDING 1] was replaced with the handshape of the CL representing the train, and the participants could understand that the person arrived by train instead of by plane.

Table 2 lists the average accuracy and realism scores for each JSL sentence. Note that, the scores in Table 2 are not limited to the CL predicate part of the JSL sentence shown in Figure 11, but are scores for the entire JSL sentence. Regarding both accuracy and realism, there was no significant difference between whether or not the sentence had

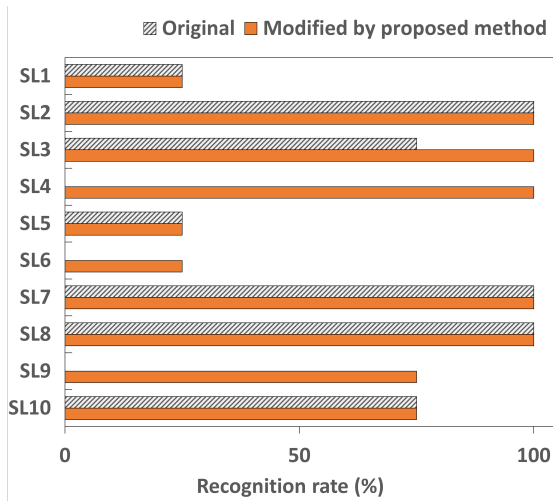


Figure 11: Recognition rate of the CL predicate part for each JSL sentence.

been modified by our proposed method. In other words, it was shown that by editing the MoCap data modified by the proposed method, contextual modifications can be reproduced without adversely affecting the quality of the data. Regarding SL10, there was no difference in recognition rate between original and modified sentence, but both accuracy and realism were improved. This is presumably due to the effect of modifying the handshape for JSL, which was expressed in the translation result as the handshape representing the CL of the car, to a handshape representing the CL of the bicycle to match the context. Also, interviews with participants revealed that three out of four were able to understand from the context that the car’s CL was incorrect in the original video during the experiment.

SL	Accuracy		Realism	
	Original	Modified	Original	Modified
SL1	2.50	2.75	3.25	3.00
SL2	3.50	3.50	3.75	3.50
SL3	3.0	2.75	3.00	2.75
SL4	3.75	3.00	3.75	3.50
SL5	3.50	3.00	3.50	3.50
SL6	3.50	3.00	3.00	3.00
SL7	3.50	3.50	3.00	3.50
SL8	3.00	3.50	3.25	3.50
SL9	3.50	3.00	2.75	3.25
SL10	2.75	3.25	2.75	3.25
Mean	3.25	3.13	3.20	3.28

Table 2: Average accuracy and realism scores for each JSL sentence.

These evaluation results demonstrate that our method achieves the flexibility to generate contextual modifications and new movements while

preserving the quality of natural human-like movements without the need for additional MoCap processes. In our experiment, there were no significant differences in the evaluation results between the Deaf, hard of hearing, and CODA persons, but as a future challenge, we need to confirm the reproducibility of the proposed method’s effectiveness by increasing the number of participants.

## 6. Conclusion

In this paper, we presented our HamNoSys-based sign language motion data editing method. This method is a hybrid that combines two utterance synthesis methods: a MoCap data-driven method and a phonological-based generation method. We implemented this method in the motion editing tool of our JSL translation system and confirmed that it is possible to edit the citation form of signs included in the JSL gloss string of the translation results as CL predicates. Our evaluation experiment revealed that by applying motion modification to the translation results using the proposed method, the intelligibility of the JSL avatar animation was improved. The proposed method achieved both natural human-like movements and the flexibility to generate contextual modifications and new movements without any additional MoCap processes. Additionally, since HamNoSys supports the transcription of all sign languages, it is also possible to create motions for other sign languages by using our JSL motion data.

In future work, we plan to investigate ways of supporting other contextual modifications such as directional verbs by considering the action component of HamNoSys. We will also explore supporting NMMs such as facial expressions and mouth gestures, which are semantically important components of SL.

## 7. Acknowledgements

We wish to thank Thomas Hanke, Maria Kopf and the researchers at the Institute of German Sign Language and Communication of the Deaf at the University of Hamburg for their valuable inputs and assistance with the explanation of HamNoSys provided in our work. We would also like to thank the members of the Intelligent Automatic Sign Language Translation (EASIER) project for their valuable suggestions and sharing information about their latest research.



## 8. Bibliographical References

- Andreas Aristidou and Joan Lasenby. 2011. Fabrik: A fast, iterative solver for the inverse kinematics problem. *Graphical Models*, 73(5):243–260.
- Heike Brock and Kazuhiro Nakadai. 2018. [Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese Sign Language expressions](#). In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4247–4252, Miyazaki, Japan. European Language Resources Association (ELRA).
- Sarah Ebling. 2016. Automatic translation from german to synthesized swiss german sign language. *PhD Thesis (University of Zurich)*.
- Ralph Elliott, John R. W. Glauert, Richard Kennaway, Ian Marshall, and Éva Sáfár. 2007. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.
- Michael Filhol and John C. McDonald. 2018. [Extending the AZee-Paula shortcuts to enable natural proform synthesis](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 45–52, Miyazaki, Japan. European Language Resources Association (ELRA).
- Michael Filhol and John C. McDonald. 2020. [The synthesis of complex shape deployments in sign language](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 61–68, Marseille, France. European Language Resources Association (ELRA).
- Stavroula-Evita Fotinea, Eleni Efthimiou, George Caridakis, and Kostas Karpouzis. 2007. A knowledge-based sign synthesis architecture. *Universal Access in the Information Society*, 6(4):405–418.
- Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The signcom system for data-driven animation of interactive virtual signers: methodology and evaluation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(1):1–23.
- Thomas Hanke. 2004. [HamNoSys – representing sign language data in language resources and language processing contexts](#). In *Proceedings of the LREC2004 Workshop on the Representation and Processing of Sign Languages: From Sign-Writing to Image Processing. Information techniques and their implications for teaching, documentation and communication*, pages 1–6, Lisbon, Portugal. European Language Resources Association (ELRA).
- Matt Huenerfauth. 2006. Generating american sign language classifier predicates for english-to-asl machine translation. *PhD thesis (University of Pennsylvania)*.
- Matt Huenerfauth, Pengfei Lu, and Hernisa Kacorri. 2015. Synthesizing and evaluating animations of american sign language verbs modeled from motion-capture data. *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2015)*, pages 22–28.
- Hiroyuki Kaneko, Narichika Hamaguchi, Mamoru Doke, and Seiki Inoue. 2010. Sign language animation using tvml. *Proceedings of the 9th ACM SIGGRAPH Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI '10)*, pages 289–292.
- Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee, and Jong Park. 2022. [Sign language production with avatar layering: A critical use case over rare words](#). In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 1519–1528, Marseille, France. European Language Resources Association (ELRA).
- John C. McDonald and Michael Filhol. 2021. Natural synthesis of productive forms from structured descriptions of sign language. *Machine Translation*, 35(4):1–24.
- John C. McDonald, Rosalee Wolfe, Jerry Schnepf, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of american sign language. *Universal Access in the Information Society*, 15(4):551–566.
- Taro Miyazaki, Tsubasa Uchida, Naoki Nakatani, Hiroyuki Kaneko, and Masanori Sano. 2023. Machine translation to sign language using post-translation replacement without placeholders. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*.
- Sara Morrissey. 2008. Data-driven machine translation for sign languages. *PhD Thesis (Dublin City University)*.

- Lucie Naert. 2020. Capture, annotation and synthesis of motions for the data-driven animation of sign language avatars. *PhD Thesis (Southern Brittany University)*.
- Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2020. [LSF-ANIMAL: A motion capture corpus in French Sign Language designed for the animation of signing avatars](#). In *12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6008–6017, Marseille, France. European Language Resources Association (ELRA).
- Lucie Naert, Caroline Larboulette, and Sylvie Gibet. 2021. Motion synthesis and editing for the generation of new sign language content: Building new signs with phonological recombination. *Machine Translation*, 35(3):405–430.
- Fabrizio Nunnari, Michael Filhol, and Alexis Heloir. 2018. [Animating AZee descriptions using off-the-shelf IK solvers](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 155–162, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020. Progressive transformers for end-to-end sign language production. *European Conference on Computer Vision (ECCV)*, pages 687–705.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5141–5151.
- William C. Stokoe, Dorothy C. Casterline, and Carl G. Croneberg. 1965. *A dictionary of American sign language on linguistic principles*. Gallaudet College Press.
- Tsubasa Uchida, Naoki Nakatani, Taro Miyazaki, Hiroyuki Kaneko, and Masanori Sano. 2023. Motion editing tool for reproducing grammatical elements of japanese sign language avatar animation. *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*.
- Unity Technologies. Accessed: 2024-02-29. <https://unity.com/>.
- Rosalee Wolfe, John C. McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: A question of representation. *Information*, 13(4):206.
- Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitris Metaxas. 2022. [Sign language video anonymization](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 202–211, Marseille, France. European Language Resources Association (ELRA).

# SignaMed: a Cooperative Bilingual LSE-Spanish Dictionary in the Healthcare Domain

Manuel Vázquez-Enríquez\* , José Luis Alba-Castro\* , Ania Pérez-Pérez\*<sup>†</sup>,  
Carmen Cabeza-Pereiro<sup>†</sup> , Laura Docío-Fernández\* 

\* AtlanTTic Research Center, University of Vigo,

<sup>†</sup> Translation and Linguistic Department,  
Campus Universitario de Vigo, Spain

\*{mvazquez, jalba, aperez, ldocio}@gts.uvigo.es

<sup>†</sup>cabeza@uvigo.es

## Abstract

In this paper we present SignaMed, a bilingual dictionary accessible in Spanish and LSE (Spanish Sign Language) specific to the medical domain. Building a sign language dataset to develop machine learning algorithms and linguistic studies is a complex task that requires the cooperation of Deaf people. The dictionary platform, built with their contributions, offers diverse access modes for users, including basic search functionalities, games, and activities for sign donation. It allows sign searching using webcam or mobile phone capturing, facilitating intuitive interaction and feedback. The article presents the technical, linguistic and cooperation details behind the construction of the dictionary and will hopefully serve as inspiration for similar initiatives in other sign languages. The dictionary is accessible through <https://signed.web.app>.

**Keywords:** LSE, Dictionary, Sign recognition, Deaf collaboration

## 1. Introduction

The landscape of sign language dictionaries is broad and diverse, driven by the intrinsic need of educators and relatives of Deaf individuals to learn sign language for communication. It is essential to remember that dictionaries play a crucial role in the consolidation of a national language, which includes sign languages. Many languages have dictionaries and sign banks collected by one or more entities, usually accessible online, where users can search for signs by keyword and view video recordings of the signs. Examples include ASL ([www.signasl.org](http://www.signasl.org)), BSL ([www.signbsl.com](http://www.signbsl.com), [bslsignbank.ucl.ac.uk](http://bslsignbank.ucl.ac.uk)), DGS ([web.dgs-korpus.de](http://web.dgs-korpus.de)), AUSLAN ([auslan.org.au](http://auslan.org.au)), LSE ([fundacioncnse-dilse.org](http://fundacioncnse-dilse.org)), NZSL ([www.nzsl.nz](http://www.nzsl.nz)), LSFB ([dicto.lsfb.be](http://dicto.lsfb.be)), among others. The European initiative Spreadthe-sign ([www.spreadthesign.com](http://www.spreadthesign.com)) is notable for compiling signs in multiple languages for comparison.

Traditionally, sign language dictionaries have not been used to train automatic recognition algorithms for several reasons: there is usually no more than one sample per sign, there are usually few signers, and because they contain isolated signs, there is a lack of information about the non-manual components and co-articulation effects. But recently they have started to be used to obtain visual references to train sign spotting algorithms that help to look up examples of the dictionaries in videos with continuous signing (Jiang et al., 2021; Varol et

al., 2022; Vázquez Enríquez et al., 2023), which opens the door to dense annotation of continuous SL footage, to advance in the translation problem, and to develop actual applications for search and retrieval.

Despite the advancements on the performance of sign spotting and isolated sign language recognition (ISLR), there has been very few examples of sign recognition models applied to practical use cases. One of the oldest examples can be found in Muhammed et al. (2016), where the authors introduced an interactive platform for communicating with Deaf individuals in a hospital setting through directed dialogue and a recognizer capable of identifying 33 signs using Dynamic Time Warping-based classifiers on RGB+D inputs from KinectV2. More recently, deep learning approaches have been utilized in small-scale applications, such as in Zhou et al. (2020), where a dataset of 45 Hong Kong Signs was collected to train a ResNet model and develop a mobile application paired with a Jetson Nano. During inference, the smartphone preprocesses the sign video, which is then wirelessly transmitted to the Jetson Nano for recognition and translation of the sign to spoken language. In the Greek project SL-ReDu, an education platform for learning GSL and providing automatic assessment (Papadimitriou et al., 2023), the authors train and test several deep learning approaches to recognize a set of 54 signs and the 24 Greek letters in fingerspelled words. They reported 91% sign recognition rate and 65% in fingerspelled letters both in signer

independent mode and using 2DCNN RGB features with a Mobilenet backbone and a BiLSTM recursive model. From a business model point of view, it seems that some start-ups are starting to leverage ISLR models, like SLAIT (now focused on an ASL educational interactive platform) or CSLR, like SignAll and Sign-Speak (still landing pages that promise ASL translation).

In this paper we detail the construction of a dictionary that allows sign lookup using isolated sign recognition algorithms as an extension of a preliminary version presented at the GoodIT2021 conference (Vázquez Enríquez et al., 2021a). To our knowledge, only a similar idea was developed simultaneously for the French-speaking sign language of Belgium (LSBF) (Jérôme et al., 2023). Their model is able to classify 700 signs with a top-10 accuracy of 83%, and responds to a query in less than 10 secs without using GPU. It is clear that bigger efforts should be made to increase the accuracy and responsiveness of these applications.

The rest of this paper is organized as follows. Section 2 presents the project and summarizes the origin, the iterative growing process and the engagement of the Deaf community. Section 3 describes the main functionalities of the dictionary platform and how volunteers can contribute. Section 4 gives some more detail on the main technology modules of SignaMed: the platform itself, the sign recognition algorithm and the quality checking for incorporating new sign donations. Section 5 is dedicated to the linguistic issues that appear when trying to build a sign language dictionary, namely the variants of signs for the same meaning and the selection and definition of LSE terms for the health domain. The paper concludes with a discussion of potential benefits and next steps for the SignaMed platform.

## 2. SignaMed: a Bilingual LSE-Spanish Dictionary

### 2.1. Origin

SignaMed was conceived from the convergence of needs during a research project on automatic recognition of Spanish Sign Language (LSE). Before the COVID-19 pandemic, we began recording a dataset of isolated signs and short phrases in a laboratory setting and at Deaf associations (Docío-Fernández et al., 2020). In that project, we surveyed the Deaf community to identify the most urgent application scenarios for deploying a potential LSE to Spanish translation service. Healthcare emerged as the top priority by a significant margin. With the pandemic making it impossible to continue sample collection in the lab and associations, we aimed to develop an online capture platform, ask-

ing the Deaf community to record health-related signs. Aware of their fatigue from long-promised technological solutions, we sought to develop a practical application using the recordings so they could immediately see their efforts were not in vain. The initial reactions to being able to search for a sign by performing it in front of a webcam or smartphone encouraged us to further invest in the platform we named SignaMed. Moreover, the medical vocabulary sparked interesting discussions about the genesis of signs in this field and the lack of signs for relatively common concepts.

The medical environment is particularly sensitive for communication. For the Deaf, it proved especially exclusionary during the Covid-19 pandemic due to mask mandates. Beyond this context, the need persists for tools that facilitate understanding between healthcare personnel and the Deaf, encouraging sign language learning at beginner levels.

SignaMed aims to break down the barriers Deaf people face with medical nomenclature and help them gain spaces of trust and privacy, which is essential for managing terminology in their own language. Healthcare personnel will become more effective with a linguistic and technological tool that enables them to explore diagnoses and name symptoms, diseases, tests, and treatments. A micro-learning course with a Telegram bot [@signasalud] was created for medical staff to learn the most relevant signs within a few weeks, enhancing communication within their environment.

### 2.2. Internal structure

SignaMed is organized according to a double search function: from LSE and from Spanish. What connects both interfaces is a system that relates signs and variants of signs with meanings or concepts (meaning labels), which correspond to a singled out definition. Internally, each variant is identified by an id-gloss, which refers to a standardized articulation, that is, it unambiguously identifies a single sign or variant. These glosses are not shared with users but used internally.

The concept of "lexical entry", traditional in lexicography, is not adequate to describe the structure of SignaMed, since the dictionary is not organized by LSE lemmas, but by signs or sets of signs associated with a concept (a meaning label). This concept is materialized in a Spanish word in the text search.

### 2.3. Growing the dictionary

The initial model for sign recognition was trained with 40 signs. It was gradually expanded through an iterative process involving the collaboration of the Federation of Associations of Deaf People of



Galicia (FAXPG), which records reference signs from the vocabulary and some common variants, and the research team, which integrates the vocabulary and videos into the dictionary. They also seek community collaboration to record new samples of the vocabulary and propose new variants that might be less common. New samples of sufficient quality are added to the training dataset for the sign recognition algorithms, and the updated model is deployed, marking new signs in the dictionary as accessible in LSE. As of this article's submission, SignaMed consists of 373 reference signs corresponding to 312 health terms from which 273 are already learnt by the model (accessible in LSE)<sup>1</sup>. SignaMed includes 273 definitions in LSE and 120 usage examples. The creation of ad-hoc definitions for the dictionary is a complex linguistic exercise, noteworthy because medical term definitions in LSE are scarce. Claudia Domínguez, a Deaf person with a master's in Applied Linguistics, first developed the definitions in Spanish, so that they were easily translatable in LSE, consulting multiple sources of Spanish definitions. Then, she translated them to LSE thus ensuring full accessibility for Deaf users seeking to understand terms in their native language. After the definitions are prepared in LSE, with the necessary adaptations, the Spanish versions are not revised.

The iterative process of constructing SignaMed is summarized in Figure 1.

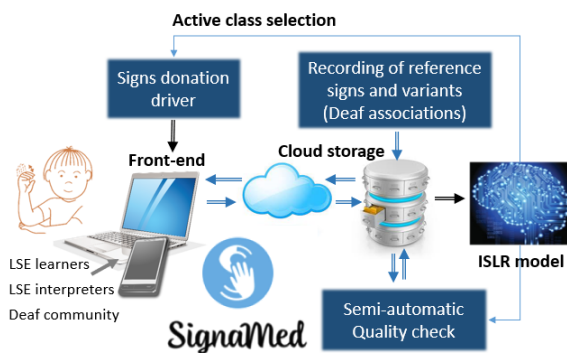


Figure 1: Iterative process for growing the SignaMed dictionary in signs and model capabilities.

The front-end allows target users to interact with the dictionary by searching words/signs and contributing with video donations. Reference signs and variants are provided by FAXPG, iteratively, including more and more specialized meanings.

<sup>1</sup>As of February 2024 the model was trained with 6K curated donations of 273 signs, but there are already more than 2K donations and 38 new signs ready to be processed for a new model. The donated video samples will not be shared due to GDPR restrictions but their Mediapipe keypoints will be made publicly available soon.

The ISLR model is trained from the dataset formed by reference signs and donations through a semi-supervised loop that curates samples and asks the users to donate specific class samples. The platform is engineered to request samples of signs most needed to enhance the model's capabilities. It's well understood that as the number of classes increases, the performance of multiclass classifiers decreases. This platform employs Active Class Selection techniques (Bicego et al., 2023) to prioritize the signs (Classes) the model needs to recognize better, whether due to insufficient samples in prior training, the shifting of decision boundaries after adding more classes, or the multiclass model partitioning the space differently in the latest growth iteration. A module named "Donate Signs" has been implemented, prompting donors to perform a series of signs requested by the system to fulfill its learning needs. Users can donate signs in this manner or contribute a new term, an unconsidered variant, or simply an additional repetition during any dictionary query.

Unfortunately, the long-term growth of the dictionary is not guaranteed, as it is being built with intermittent public funding, but the research groups involved are firmly committed to making the application increasingly useful, both for research and for everyday use, by searching alternative funding options.

## 2.4. Engaging the Deaf Community

Engaging the Deaf community in today's vast landscape of mobile applications is a challenge, which has led to the creation of a collaborative project that involves potential users of the application in its creation, incorporating playful and educational activities related to the underlying technology. SignaMed emerges as a citizen science project in which the Deaf community acts as both contributor and beneficiary. This approach requires maintaining optimal usability and drawing attention to the functionality of the application, ensuring that users not only understand its fundamentals, but also to comprehend how the machine makes use of generalizations about movement that exclude the reuse of the personal image and thus ensure anonymity.

A dedicated website<sup>2</sup> features videos in LSE explaining critical aspects of the algorithms for extracting spatial-temporal features defining signs and their classification, emphasizing personal data privacy and management within the SignaMed platform. Additionally, the platform offers interactive activities to highlight the importance of recording quality using webcams or mobile phones for the dictionary search. Users can compete for the highest scores by correctly identifying signs based on

<sup>2</sup>[www.signamed.uvigo.es](http://www.signamed.uvigo.es)

movement and articulations, with varying recording quality, and vie for the best quality recordings as a personal challenge, thereby enriching the platform with high-quality signs for continued growth, as illustrated in Figure 1.

### 3. Main Functionalities

With the development and evolution of the application, new features were added to the basic dictionary functions to encourage participation and interest from both the Deaf community and those interested in the field, thus promoting knowledge and collaboration within the platform.

Upon accessing SignaMed, the initial window (Figure 2) presents various tools and activities available to users. From left to right, these include video search, text search, "Donate Sign," and some games to explain the technology while playing.

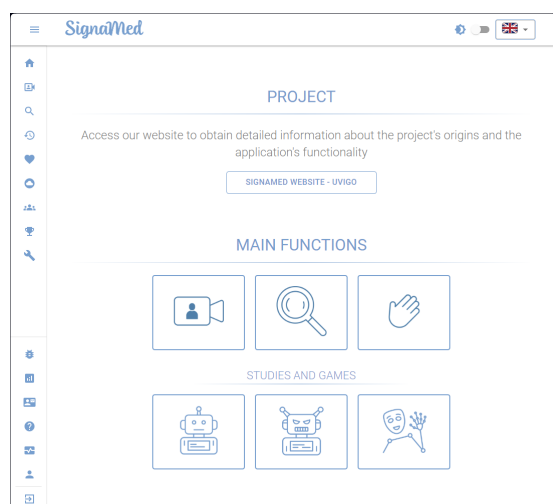


Figure 2: Home screen for the SignaMed platform

The web application offers several access modes adapted to user roles, allowing for different functionalities. Guests have restricted access to the core dictionary search functions, including both text-based and video-based searches. Registered users can participate in games and other activities such as the donation of elicited signs. Annotators, expert LSE collaborators, have access to an exclusive tool for the review and validation of videos.

The Guest option is needed to allow searching the dictionary without the platform saving the video of the query sign. When someone registers is giving permission, following the EU GDPR, to save their query for the only purpose of improving the recognition model.

The most unique feature of this dictionary is the search for a sign using the webcam or the cell phone. Figure 3 shows the recording dialog. After recording, the users are prompted to verify if they

want to send it. Then the keypoints are extracted with Mediapipe Holistic (Lugaresi et al., 2019) and the keypoint matrix is injected into the trained recognition system, based on a MSG3D architecture as explained in Vázquez Enríquez et al. (2021b). Then, the user is shown the top-3 signs with their corresponding recognition confidence. In the example in Figure 3, the DIABETES term sign is recognized. The dialog allows the user to give feedback on the recognition result and even to indicate, in case the correct result is not among the top 3, which sign was asked for. In addition, the definition in sign language and an example of use in a medical environment can be consulted.

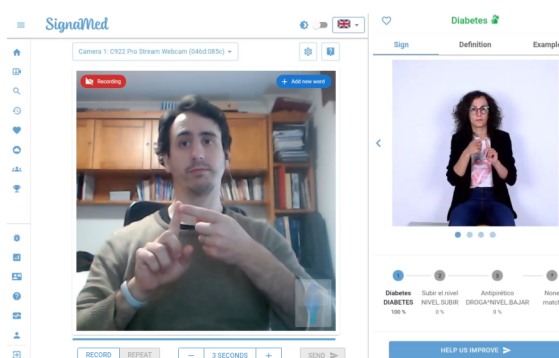


Figure 3: Dialog for sending a video query for a sign (left) and the top-3 results with their associated confidences (right)

The main functionality is common to Guests and Registered users, but the later can also donate signs. They have several options to do it:

- Looking for "red tagged" signs in the dictionary: red means that the model doesn't have enough samples for that sign to produce an accurate estimation (Figure 4 left part).
- Adding a sign variant for the same meaning: useful if the user knows another way to sign the same meaning, so they are invited to add it to the "puzzle" of variants (Figure 4 right part).
- Donate signs in a series: the users sit, relax and wait for the system to elicit the signs it needs more, so they just repeat and send until they decide to stop.

Videos from registered contributors are curated in a semiautomatic process that is explained in subsection 4.2.1.

As of December 1, 2023 SignaMed had 7050 donated signs from 339 registered users, 156 of whom have participated in the proposed interactive games. Figure 5 shows the evolution of donated signs since the first version of SignaMed. The peaks in this graph coincide with the dates

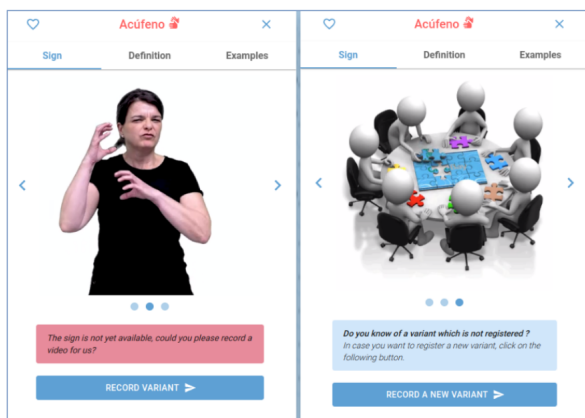


Figure 4: Option for donating a red-tagged sign (left) or a variant with the same meaning (right)

when campaigns were carried out through social networks or by going to deaf associations.

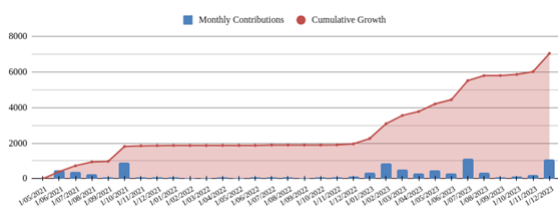


Figure 5: Evolution of sign donations in SignaMed

## 4. Technology Behind the Curtain

We summarize here the three main technical developments under the platform: the technology associated with the deployment of the platform itself, the technology that allows the recognition of signs and the semi-automatic module to check the quality of the donated signs.

### 4.1. Technology for the Deployment of the Web Application

The technologies implemented for the deployment of SignaMed were designed to provide an optimal experience on desktop browsers and mobile devices, and to safeguard the videos and data generated from user participation.

For deployment, we integrated Firebase for hosting, authentication and as a database and storage for reference videos. Cloudflare supports communication with the server, improving loading speed and offering protection against DDoS attacks. In the server, the requests pass again through several layers of security (firewall and a Nginx reverse proxy) reaching a Restful API that allows us to answer queries to our database, record videos and user activity, and process videos.

### 4.2. Sign Recognition

Figure 6 shows a summary process for automatic recognition of a query signal. We have adopted a recognition model based on keypoints (Mediapipe holistic (Lugaresi et al., 2019) in this case) because i) the sparsity of RGB samples do not guarantee a robust video-based deep neural network, and ii) when running Mediapipe in the client, the keypoint matrices weigh much less to transfer across the client-server platform which makes the whole system lighter and allows for more agile dictionary lookups.

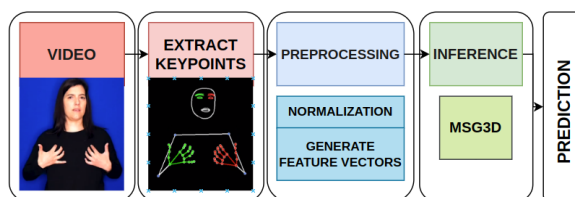


Figure 6: Sign recognition pipeline

Following the successful performance of the MSG3D-based solution merging logits of joints (keypoints) and bones (natural connections between joints) in Vázquez Enríquez et al. (2021b) we decided to train this model for the SignaMed dictionary using the samples donated by the users. The model is retrained periodically when a new set of curated signs is available.

#### 4.2.1. Quality check of the donated signs

One of the challenges of training a model when few samples are available consist of dealing with the problem of noisy data. In the SignaMed iterative process for growing the dictionary there's a necessity of cleaning the donated samples due to two main issues: videos are captured in the wild and signs might not correspond to the elicited ground-truth. These two problems were tackled with a three-stage quality check:

1. A computer vision routine automatically checks several sources of quality degradation that could hamper the correct extraction of keypoints: hands blurriness, person too close or too far from the camera, arms-hands partially missed during the sign recording, too dark or bright illumination. A score is given to each video and those with low scores are discarded in the new training set.
2. The donated samples that correspond to repetitions of signs already accessible through the model, are passed through it to check if the predicted sign corresponds to the elicited one (ground-truth). If the difference over the

second passes a safety threshold the video is included in the new training set but not tagged for manual review.

- The set of videos not discarded because of quality and not being safely classified by the current model, go through a manual review of the labels. This process is done by research group members, Deaf and hearing persons using an ad-hoc module (Figure 7) that allows reviewing around 300 samples per hour from any internet-connected device. This module allows to add comments from a predefined list regarding the video quality and the realization of the elicited sign. These annotations are very useful to improve the computer vision routine for automatic quality labeling.

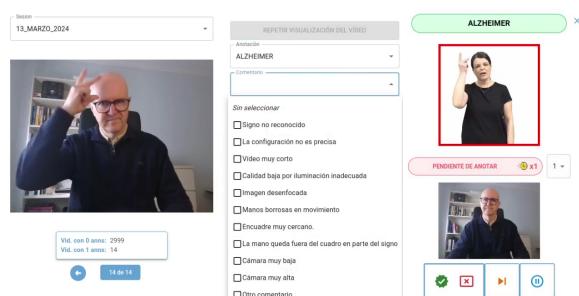


Figure 7: Module to review labels online with the quick review option (right) and the tool for adding comments if needed (left).

SignaMed does not give any instructions for donating signs, leaving freedom for each person to sign as they usually do. It is worth noting that differences have been detected in the way native speakers or interpreters donate signs, compared to people who are learning LSE. The latter group tends to imitate the sign as they see it in the video, which detracts from the naturalness of the samples. However, we have decided to keep all the videos with correct signing in order to have more samples when training the algorithms.

#### 4.2.2. Recognition Accuracy

Currently, the model is trained for recognizing 273 signs, a number continuously growing based on the availability of new curated videos from donations. The current overall performance of the model is summarized in Table 1 for the test set of reference signs (not used in training). The server responds within 3 to 4 seconds after the user submits a video, depending on its duration. This time is shortened to 350ms by extracting the keypoints directly in the browser if the user's device is able to run the Mediapipe keypoints estimator at least at 10 fps.

Stream	Top1	Top5
Joints + Bones	92	97

Table 1: Top-1 and Top-5 accuracy (%)

## 5. Linguistic Challenges in SignaMed

The consolidation of a dictionary of technical terms for a minority language is already a challenge in itself, due to the proliferation of variants that arise for the same concepts and the need to adapt word formation procedures that are natural and usable. The deaf community must be involved in this task but when facing the creation of SignaMed it is necessary to be aware of the doubts and difficulties it poses, both for Deaf individuals and for organizational entities<sup>3</sup>. Proposals for the creation of a particular term may arise simultaneously in different geographical contexts, with great insecurity on the part of its creators due to the absence of a standard. As far as the formation of new terms, the usefulness of the composition procedure has been detected (at least in the case of the LSE and for the field of health). Compound signs such as *DOC-TOR+OPERATE* (*surgeon*) are common. However, although faithful to the meaning, they are difficult to remember and constitute a challenge for automatic recognition. In the case of LSE, there are some lexicographic repertoires that constitute good sources for medical signs: [Ferre \(2006\)](#); [CNSE \(2019\)](#); [Aroca et al. \(2003\)](#).

The selection of terms and the elaboration of definitions constitute another difficulty. Definitions have to be clear, adapted to the meaning and simple. The existing lexicographic sources, both general and specialized, of spoken languages do not always constitute appropriate models. This is partly due to intrinsic features of LSE (and other sign languages), such as categorical indeterminacy, which often makes complex the exclusion of the defined term in the definition. Thus, for example, "vivir" (*to live*), "vivo" (*alive*) and "vida" (*life*) in Spanish are a single sign in LSE. Something similar happens with the polysemy of the signs: "hígado" (*liver*) and "hepatitis" (*hepatitis*) have the same sign (examples from [Domínguez, 2023](#)). In practice, this has led to ad hoc solutions, such as using a circumlocution to define hepatitis: "Inflammation of the organ that regulates the chemical levels of the blood" [Domínguez \(2023\)](#).

<sup>3</sup>In the case of LSE, there is an entity whose mission is to standardize and protect the language: the Centro de Normalización Lingüística de la LSE (CNLSE).



## 5.1. Variants of Signs in the Health Domain

As already mentioned, relatively frequently (34%) more than one sign appears associated with the same meaning label. For example, for "allergy" we recorded two different articulations, glossed as ALERGIA and ALERGIA2 (Spanish form for ALLERGY and ALLERGY2, see Figure 8).

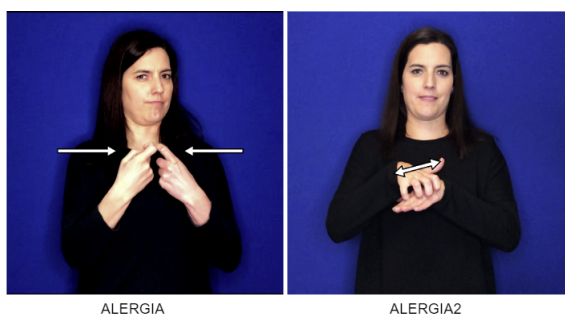


Figure 8: Two sign variants for the meaning "Allergy"

Three types of variants have been recorded:

- **Phonological:** only one parameter varies. Thus, for example, we have recorded three different articulations for "alta" (*discharge*): ALTA, ALTA(MP) and ALTA(2M). In all three the dominant hand is raised with the palm upwards, the difference lies in the passive hand: it does not intervene, it intervenes statically or it intervenes with the same movement and orientation of the dominant hand (Figure 9). In total there are 61 articulations, which are grouped into 29 meaning labels. Other examples of labels that gather phonological variants are: "análisis de sangre", "azúcar", "meningitis" or "tensión" (*blood test, sugar, meningitis or tension*, respectively).
- **Morphological:** in some cases, articulations referring to the same lemma have been recorded. These are directional verbs, whose realization is noted in different orientations "ayudar", "revisar" (*help, look-over*) or signs with relevant location like "herida" (*wound*). They represent a total of 12 signs in the database, which are grouped into 5 meaning labels.
- **Lexical:** these are the most frequent and the ones that constitute true variants. 159 signs are involved in this type of variation, grouped in 71 meaning labels. In addition to "alergia" (*allergy*), other meanings that group lexical variants are, for example: "colesterol", "diabetes", "diarrea" or "ictus" (*cholesterol, diabetes, diarrhea or ictus*, respectively).

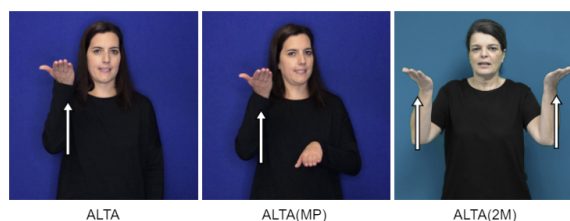


Figure 9: Three articulations for "discharge"

Two pie-charts are presented in Figure 10. The top one (signs) shows the percentage of variants, according to the types presented above. The "forms 0" include those with no registered variants and those considered reference forms<sup>4</sup>. It shows that variants constitute slightly more than a third of the total SignaMed database. The bottom one (meanings) presents a summary of the meaning labels. It focuses in how variants are grouped in relation to meanings.

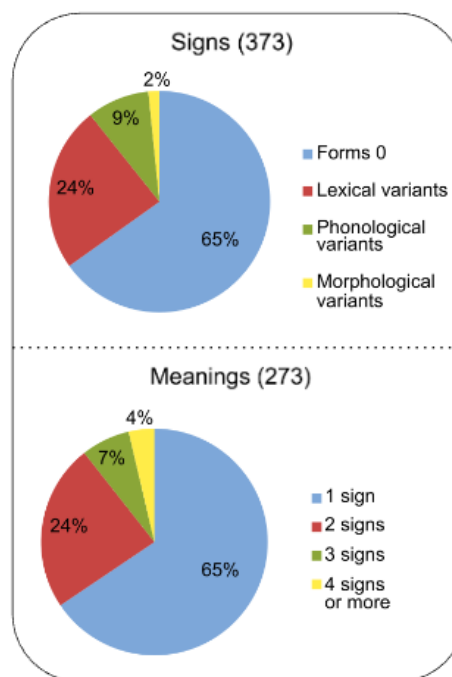


Figure 10: Distribution of variants in SignaMed

As mentioned above, about one third of the meanings into which SignaMed signs are grouped have more than one associated sign. Since the terminology tends to be univocal, one could hypothesize that, as the dictionary grows in number of signs and meanings, these groupings into variants will become less and less frequent. However, there is no indication that this will be the case. On the contrary, it is possible that variants of some of the

<sup>4</sup>Only in order to make visible in how many cases there is more than one form for the same meaning label. It is not intended to select one variant as the main one.

meaning labels that are not currently registered may appear in the future. Thus, for example, there is a single sign for "depresión" (*depression*), but "antidepresivo" (*antidepressant*) is linked with two compound signs, the second element of which is the sign for "contra" (*against*) and the first is in one case the same sign for "depresión" and the second is another form for the same meaning (the signer accompanies it with a mouthing that corresponds to the Spanish word "depresión"). The reasons that can be given for this variation are two: on the one hand, the fact that LSE is, like other sign languages, a minority and poorly standardized language. On the other hand, lexical creation procedures in sign languages have a conceptual basis strongly rooted in bodily perceptions, which is especially profitable in the case of diseases, symptoms, treatments and other semantic categories that are part of medical terminology and health.

## 5.2. Challenges in selecting terms and developing definitions

As has already been noted, one of the problems that have arisen when developing definitions is that of avoiding the defined term. For example, "hígado" (*liver*) in "hepatitis" or "pulmón" (*lung*) in "neumonía" *pneumonia*. They have been resolved with paraphrases and circumlocutions, but also by exploiting the iconic resources of the LSE. The solution of finding synonyms leads to another problem: that of deciding whether said synonymous signs are not actually lexical variants with the same meaning. Another difficulty that had to be overcome is the coincidence of the name, in Spanish, of the disease and the agent that causes it. This is what happens with "hongo" (*fungus*). In this case it was decided to provide two different entries in the dictionary. For the disease, four variants were identified, two of which begin with the fingerspelled H (in one of them followed by the sign "célula" (*cell*) and another two locating the sign for *spot* in different body places (on the arm and on the torso). For the agent that causes the disease, a compound was formed with the sign used for *mushroom* (a common and well-known type of fungus) and another glossed as *etc*<sup>5</sup>. The LSE definition proposed for the disease begins by specifying that it affects the skin tissues and then points out different locations. For the living being, a description of its characteristics and ways of life is provided. The collaborating team of the FAXPG, who was hired by the project to record signs and definitions that were being selected (see section 2.3), intervened in these decisions. The fact that LSE allows different body locations to be selected to indicate where an illness is located has also posed

<sup>5</sup>The Spanish signs corresponding to the meanings *spot*, *mushroom* and *etc* are not searchable in SignaMed.

some difficulty. In the case of "infarto" (*infarction*) there is a generic sign that does not specify a location. Due to this, a generic entry has been included in the dictionary, another for "infarto de miocardio" (*myocardial infarction*) (whose sign consists of a compound whose first part indicates the location of the heart and the second is "INFARTO") and a third for "ictus". The latter has five variants, one of which is a compound in which the first term points to the head and the second is "INFARTO".

## 6. Concluding remarks and next steps

In this article we have presented SignaMed, an accessible collaborative bilingual LSE-Spanish dictionary in the health domain. The dictionary is conceived as a citizen science project to involve its recipients in the process of building and learning the AI techniques that support it. The article is intended to serve as an example of the necessary collaboration that must exist in any project that seeks to develop sign recognition or sign language translation technology. Brief details of each of the main parts of the project have been given, but due to space limitations some functionalities have been left out. The reader is invited to try it out at <https://signamed.web.app>.

The next steps for the SignaMed platform are already underway: preparing it for extension to translation of phrases in the healthcare domain. The challenge is to get the Deaf community to contribute phrases for a specific purpose. The platform is already preparing to learn a communication ontology in a hospital emergency department where there is an established protocol of questions. The SignaMed platform will have all questions and samples of potential answers in LSE. Donors will be able to choose between signing exactly the same answer, some glossing variant with the same meaning, or a totally different answer. These interactions will help to tune an end2end sign language translator between LSE and Spanish in the healthcare domain.

In short, this project serves the dual purpose of demonstrating a practical use of isolated sign recognition technology while presenting a user-friendly signs collection platform that can be used for new projects.

## 7. Acknowledgements

We would like to acknowledge the contribution of FAXPG, FCNSE, Claudia Domínguez and Manuel Lema for the recording of reference signs and definitions in LSE and the review of donated signs, as well as the contribution of anonymous collaborators for their donations of sign instances. This

work has been supported by the Spanish projects PID2021-123988OB-C32, FCT-21-16924 and by the Xunta de Galicia and ERDF through the Consolidated Strategic Group AtlanTTic (2019-2022). Manuel Vázquez Enríquez is funded by the Spanish Ministry of Science and Innovation through the predoc grant PRE2019-088146.

## 8. Bibliographical References

- E. Aroca et al. 2003. *Salud: Medicina*. Fundación CNSE.
- Manuele Bicego, Manuel Vázquez-Enríquez, and José L. Alba-Castro. 2023. Active class selection for dataset acquisition in sign language recognition. In *Image Analysis and Processing – ICIAP 2023*, pages 304–315. Springer Nature Switzerland.
- Fundación CNSE. 2019. *DILSE. Diccionario de la Lengua de Signos Española. Diccionario básico*. Fundación CNSE.
- L. Docío-Fernández et al. 2020. [LSE\\_UVIGO: A multi-source database for Spanish Sign Language recognition](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages.*, pages 45–52.
- Claudia Domínguez. 2023. *SignaMed: recurso léxico sobre la salud para personas sordas*. Master's thesis, Universidade de Vigo.
- J.M. Ferre. 2006. *Ámbito de sanidad: frases, diálogos, vocabulario: curso de Lengua de Signos Española*. Centro Altatorre de Personas Sordas.
- Tao Jiang, Necati Cihan Camgoz, and Richard Bowden. 2021. [Looking for the signs: Identifying isolated sign instances in continuous video footage](#). In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE.
- J. Jérôme et al. 2023. [Sign language-to-text dictionary with lightweight transformer models](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5968–5976. International Joint Conferences on Artificial Intelligence Organization. AI for Good.
- C. Lugaresi et al. 2019. [Mediapipe: A framework for perceiving and processing reality](#). In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*.
- S. Muhammed et al. 2016. Hospisign: An interactive sign language platform for hearing impaired. *Journal of Naval Sciences and Engineering*, 11(3):75–92.
- K. Papadimitriou et al. 2023. [Greek sign language recognition for an education platform](#). *Universal Access in the Information Society*, pages 1–18.
- G. Varol et al. 2022. Scaling up sign spotting through sign language dictionaries. *International Journal of Computer Vision (IJCV)*, pages 1416–1439.
- M. Vázquez Enríquez et al. 2021a. [Deep learning and collaborative training for reducing communication barriers with deaf people](#). In *Proc. Conf. on IT for Social Good, GoodIT '21*, page 289–292, NY, USA. ACM.
- M. Vázquez Enríquez et al. 2021b. [Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3457–3466.
- M. Vázquez Enríquez et al. 2023. [Eccv 2022 sign spotting challenge: Dataset, design and results](#). In *European Conference on Computer Vision Workshops (2022)*, pages 225–242, Cham. Springer Nature Switzerland.
- Z. Zhou et al. 2020. [A portable hong kong sign language translation platform with deep learning and jetson nano](#). In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '20*.

# Diffusion Models for Sign Language Video Anonymization

Zhaoyang Xia<sup>1</sup>, Yang Zhou<sup>1</sup>, Ligong Han<sup>1</sup>, Carol Neidle<sup>2</sup>, Dimitris N. Metaxas<sup>1</sup>

<sup>1</sup> Rutgers University, <sup>2</sup> Boston University

<sup>1</sup> 110 Frelinghuysen Road, Piscataway, NJ 08854,

<sup>2</sup> Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215

zx149@rutgers.edu, eta.yang@rutgers.edu, lh599@scarletmail.rutgers.edu,

carol@bu.edu, dnm@cs.rutgers.edu

## Abstract

Since American Sign Language (ASL) has no standard written form, Deaf signers frequently share videos in order to communicate in their native language. However, this does not preserve privacy. Since critical linguistic information is transmitted through facial expressions, the face cannot be obscured. While signers have expressed interest, for a variety of applications, in sign language video anonymization that would effectively preserve linguistic content, attempts to develop such technology have had limited success and generally require pose estimation that cannot be readily carried out in the wild. To address current limitations, our research introduces DiffSLVA, a novel methodology that uses pre-trained large-scale diffusion models for text-guided sign language video anonymization. We incorporate ControlNet, which leverages low-level image features such as HED (Holistically-Nested Edge Detection) edges, to circumvent the need for pose estimation. Additionally, we develop a specialized module to capture linguistically essential facial expressions. We then combine the above methods to achieve anonymization that preserves the essential linguistic content of the original signer. This innovative methodology makes possible, for the first time, sign language video anonymization that could be used for real-world applications, which would offer significant benefits to the Deaf and Hard-of-Hearing communities.

**Keywords:** Sign Language Anonymization, Diffusion Model, Text-to-Video Editing, ASL

## 1. Introduction

American Sign Language (ASL), the predominant language used by the Deaf Community in the US and parts of Canada, is a full-fledged natural language. It employs manual signs in parallel with non-manual elements, including facial expressions and movements of the head and upper body, to convey linguistic information. The non-manual elements are crucial for conveying many types of lexical and adverbial information, as well as for marking syntactic structures (e.g., negation, topics, question status, and clause types (Baker-Shenk, 1985; Kacorri and Huenerfauth, 2016; Neidle et al., 2000; Coulter, 1979; Valli and Lucas, 2000)). Consequently, in video communications, e.g., on the Web, involving sensitive subjects such as medical, legal, or controversial matters, obscuring the face for purposes of anonymity would result in significant loss of essential linguistic information.

Despite the fact that several writing systems have been developed for ASL (Arnold, 2009), the language has no standard written form. While ASL signers could use written English in order to preserve privacy, that is frequently not their preference, as signers generally have greater ease and fluency in their native language, ASL, than in English.

Many Deaf signers have shown interest in a mechanism that would maintain the integrity of linguistic content in ASL videos while disguising the identity

of the signer, as discussed in several recent studies (Lee et al., 2021). There are many potential applications of such a tool. For example, this could enable anonymous peer review for academic submissions in ASL. This could also ensure impartiality in various multimodal ASL-based applications, e.g., enabling production of neutral definitions for ASL dictionaries, not tied to the identity of the signer producing them. It could also enable maintenance of neutrality in interpretation scenarios. Additionally, such a tool could increase signers' willingness to contribute to video-based AI datasets (Bragg et al., 2019b), which hold significant research value.

For these reasons, privacy preservation for ASL videos has been explored (Isard, 2020). However, most of these approaches suffer from limitations with respect to preservation of linguistic meaning, and they generally achieve only a limited degree of anonymity. They also require accurate pose estimation, and some require substantial human labor. These limitations significantly reduce the potential for practical applications of such technologies.

To overcome the limitations of existing anonymization tools, we introduce DiffSLVA, a novel anonymization approach leveraging large-scale pre-trained diffusion models, notably Stable Diffusion (Rombach et al., 2022). DiffSLVA is designed to tackle text-guided sign language anonymization. Through a text prompt, it generates a new video in which the original linguistic meaning is retained, but



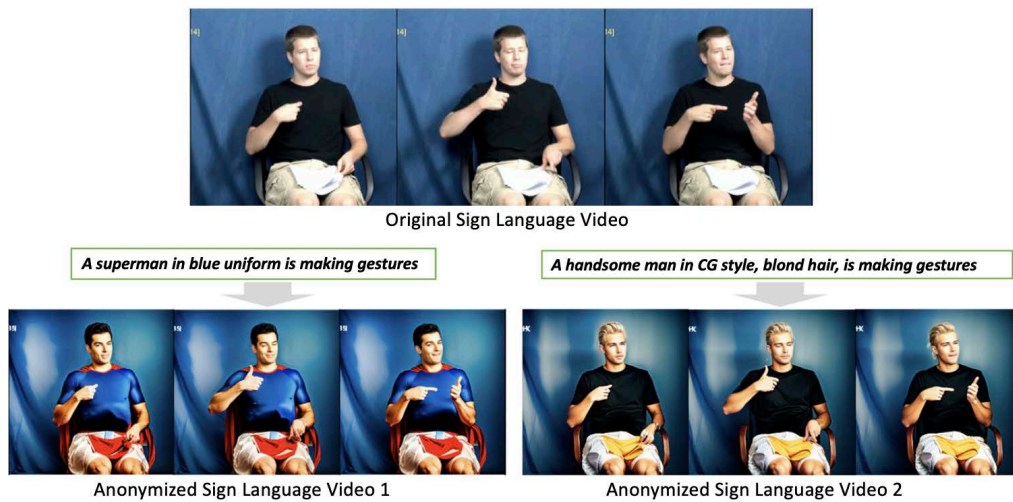


Figure 1: **Text-guided Sign Language Video Anonymization.** We introduce DiffSLVA, an innovative approach that leverages the capabilities of diffusion models to achieve text-guided sign language video anonymization. This method is capable of anonymizing sign language videos with a single text prompt, effectively masking the identity of the original signer while preserving the linguistic content and nuances.

the identity of the signer is altered. Figure 1 illustrates the method. Unlike traditional methods that require skeleton extraction, our approach uses the Stable Diffusion model enhanced with ControlNet (Zhang et al., 2023) to process language videos with Holistically-Nested Edge (HED) (Xie and Tu, 2015), which can more easily and robustly process videos in the wild. To adapt the image-based Stable Diffusion for video, we follow Yang et al. (2023), but modify the methods. We replace the self-attention layer in U-Net with a cross-frame attention layer and implement an optical-flow-guided latent fusion for consistent frame generation. Additionally, to capture fine-grained facial expressions, we have developed a specialized facial generation module using a state-of-the-art image animation model (Zhao and Zhang, 2022) fine-tuned on our mixed dataset (see Section 4.1). The outcomes are integrated via a face segmentation technique (Yu et al., 2018). Our results show substantial promise for anonymization applications, which would be invaluable for the Deaf and Hard-of-Hearing communities.

Our work makes several key contributions to the field of sign language video anonymization:

- (1) We propose text-guided sign language anonymization. The anonymized videos are based on computer-generated humans, transforming the original signer’s appearance to that of a computer-generated individual.
- (2) We have developed a specialized module dedicated to improving facial expression transformation. Our ablation studies show that this significantly enhances the preservation of linguistic meaning.
- (3) Our approach relies solely on low-level image features, such as edges, enhancing the potential for practical applications.
- (4) Our anonymization can accommodate a diverse range of target humans. The anonymized signers

can have any ethnic identity, gender, clothing, or facial style, a feature many ASL signers want; this simply requires changing the text input.

## 2. Related Work

### 2.1. Video Editing with Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2020) have shown exceptional performance in the field of generative AI. Once trained on large-scale datasets (e.g., LAION (Schuhmann et al., 2022)), text-guided latent diffusion models (Rombach et al., 2022), e.g., Stable Diffusion, are capable of producing diverse and high-quality images from a single text prompt. Additionally, ControlNet (Zhang et al., 2023) presents a novel enhancement. It fine-tunes an additional input pathway for pre-trained latent diffusion models, enabling them to process various modalities, including edges, poses, and depth maps. This innovation significantly augments the spatial control capabilities of text-guided models.

Image-based diffusion models can also be used for video generation or editing. There have been efforts to modify image-based diffusion models for consistent generation or editing across frames. Tune-A-Video (Wu et al., 2023) inflates a pre-trained image diffusion model, modified with pseudo 3D convolution and cross-frame attention and then fine-tuned on a given video sequence. During the inference stage, with the DDIM inversion noises (Song et al., 2020) as the starting point, the fine-tuned model is able to generate videos with similar motions but varied appearance. Edit-A-Video (Shin et al., 2023), Video-P2P (Liu et al., 2023), and vid2vid-zero (Wang et al., 2023) utilize Null-Text Inversion (Mokady et al., 2023) for improved reconstruction of video frames, which

provides better editing results. Fine-tuning or optimization based on one or more input video sequences is required by these methods. Moreover, the detailed motion in the video cannot be captured properly without having a negative impact on the editing abilities. Therefore, they are not suitable for the sign language video anonymization task.

Other methods use the cross-frame attention mechanism or latent fusion to achieve the video editing or generation ability of image-based diffusion models. Text2Video-Zero (Khachatryan et al., 2023) modifies the latent codes and attention layer. FateZero (Qi et al., 2023) blends the attention features based on the editing masks detected by Prompt-to-Prompt (Hertz et al., 2022). Pix2Video (Ceylan et al., 2023) aligns the latent features between frames for better consistency. Rerender-A-Video (Yang et al., 2023) utilizes a cross-frame attention mechanism and cross-frame latent fusion to improve the consistency of style, texture, and details. It can also be used with ControlNet for spatial guidance. However, these methods cannot accurately transfer facial expressions from the original videos. Therefore, they lose a significant amount of the linguistic meaning from the original video. Our approach is based on the Rerender-A-Video (Yang et al., 2023) method, without the post video processing, to best capture manual signs. To overcome the loss of linguistically important non-manual information, we designed a specialized facial expression translation module (Zhao and Zhang, 2022), which we combine with the rest of the anonymized body using a face parser model (Yu et al., 2018).

## 2.2. Sign Language Video Anonymization

Various strategies have been explored for privacy preservation in ASL video communication (Isard, 2020). Early approaches used graphical filters, such as a tiger-shaped filter (Bragg et al., 2019b), to disguise the face during signing. However, these filters often lead to a loss of critical facial expressions, thereby hindering comprehension. Alternatives like blocking parts of the face (Bleicken et al., 2016) also result in significant information loss. Approaches involving re-enacting signed messages with actors (Isard, 2020) or using virtual humans for anonymous sign language messaging (Heloir and Nunnari, 2016; Efthimiou et al., 2015) are labor-intensive, challenging, and time-consuming.

Some approaches to avatar generation for sign language, e.g., that of Bragg (2019a), use cartoon-like characters to replace signers. Cartoonized Anonymization (Tze et al., 2022b) proposes use of pose estimation models (Li et al., 2018; Xiu et al., 2018; Lugaresi et al., 2019) to automatically enable the avatars to sign. Yet, these methods often lead to unrealistic results (Kipp et al., 2011).

Deep-learning approaches, such as AnonySign (Saunders et al., 2021) or Neural Sign Reenactor (Tze et al., 2022a), leverage GAN-based methods for photo-realistic sign language anonymization using skeleton keypoints for accurate image generation. The results are encouraging. However, they require accurate skeleton keypoints and face landmarks. In sign language videos, rapid hand movements can lead to blurring in the video frames. Occlusions of the face by the hands also occur frequently. For these reasons, the performance of existing human pose estimation models is often inadequate when applied to sign language videos, which leads to errors in the anonymized video.

Recent work (Lee et al., 2021) applies the facial expression transfer method of Siarohin et al. (2019b) for sign language anonymization. This method involves replacing the signer’s face in the video with another individual’s face, while transferring the facial expressions to the new face. As a result, this approach successfully preserves the linguistic meanings conveyed by facial expressions and alters the identity of the signer in the video. However, in Lee et al. (2021), the extent of the anonymization is not complete, since only the face is replaced, while the arms, torso, and hands remain the same as in the original video. Another method (Xia et al., 2022) uses an unsupervised image animation method (Siarohin et al., 2021; Ren et al., 2020) with a high-resolution decoder and loss designed for the face and hands to transform the identity of a signer to that of another signer from the training videos. The results are promising. However, this method can work well only in the training data domain with limited signer identities and is hard to adapt to sign language videos in the wild.

To address the above limitations, we propose Diff-SLVA, a method that is based on the modification of large-scale diffusion models and ControlNet for consistent high-fidelity video generation, which can be used to achieve effective sign language video anonymization in the wild. Our approach is a text-guided sign language video anonymization, as shown in Figure 1. For the anonymization of signers’ body, arms and hands, we use large-scale diffusion models, which do not rely on the use of sign language video data for training and can perform zero-shot sign language video anonymization. With the help of ControlNet, we use low-level features instead of accurate skeleton data as signal for generation guidance, so that the results are not adversely affected by inaccurate skeleton estimations. To further improve the facial expression translation, we designed a specialized model for facial expression enhancement and combine it with the model that anonymizes the rest of the body using a face parser model. Our method can anonymize sign language videos based on a single text prompt. The

anonymized video is based only on a wide range of computer-generated humans. Our anonymization technique thereby offers great promise for applications that would benefit the Deaf community.

### 3. Methodology

In this section, we introduce our method for text-guided sign language video anonymization. The process is structured as follows: Given a sign language video with  $N$  frames  $\{I_i\}_{i=0}^N$ , we use a pre-trained latent diffusion model, augmented with ControlNet, to execute the anonymization. A text prompt  $c_p$  serves as guidance for the desired anonymization identity or style. Our goal is to generate an altered video sequence, represented by  $\{I'_i\}_{i=0}^N$ , that conceals the identity of the original signer while preserving the linguistic content.

In 3.1, we introduce the text-guided latent diffusion models and the ControlNet, which serve as the foundation for text-guided image generation. Section 3.2 details the methods for adapting the text-to-image method for consistent video editing. To ensure preservation of linguistic meaning through accurate facial expression translation, we introduce a specialized facial enhancement module in 3.3. Figure 2 shows an overview of our method.

#### 3.1. Latent Diffusion Models

Latent diffusion models operate in the latent space for faster image generation. The input image  $I$  is first input to an encoder  $\varepsilon$  to obtain its latent features  $x_0 = \varepsilon(I)$ . The following diffusion forward process adds noise to the latent features:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (1)$$

where  $t = 1, \dots, T$  is the time step indicating the level of noises added;  $q(x_t|x_{t-1})$  is the conditional probability of  $x_t$  given  $x_{t-1}$ ; and  $\alpha_t$  are hyperparameters that adjust the noise level across the time step  $t$ . Leveraging the property of Gaussian noise, we can also sample  $x_t$  at any time step by the following equation:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (2)$$

where  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ .

In the diffusion backward process, a U-Net  $\epsilon_\theta$  is trained to estimate the above added noise to recover  $x_0$  from  $x_T$ . For the conditional diffusion model,  $\epsilon_\theta$  takes the conditional information  $c_p$  as input to guide the generation process. After  $\epsilon_\theta$  has been trained, the  $x_{t-1}$  can be sampled by strategies such as DDIM sampling (Song et al., 2020):

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t, c_p), \quad (3)$$

where  $\epsilon_\theta(x_t, t, c_p)$  is the predicted noise at time step  $t$ . For the DDIM sampler, we can estimate the final

clear output  $\hat{x}_0$  at each time step  $t$ .  $\hat{x}_0$  can also be represented as the following equation:

$$\hat{x}_0 = (x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, c_p))/\sqrt{\bar{\alpha}_t}, \quad (4)$$

During inference, for a Gaussian noise  $x_T$ , we can sample a clear latent  $x_0$  with the DDIM Sampler and decode it to the generated image  $I' = D(x_0)$ . Our methodology also incorporates ControlNet, introducing an additional signal to the text-guided latent diffusion models. This structure makes it possible for the text-guided diffusion model to take diverse inputs like edges, human poses, and segmentation maps for more spatial constraints. Consequently, with incorporation of an additional input  $c_n$ , the predicted noise at each time step  $t$  is represented as  $\epsilon_\theta(x_t, t, c_p, c_n)$ . This approach enhances the alignment of the final outputs with the spatial features specified by the input condition  $c_n$ .

#### 3.2. Consistent Video Generation

Although Stable Diffusion models exhibit outstanding performance in image generation, application to videos is challenging. Directly applying Stable Diffusion to videos gives rise to significant frame inconsistency issues. To address this, we adapt text-to-image diffusion models for video editing tasks, drawing upon the framework established by Yang et al. (2023). Our approach begins by encoding and sampling the original frames  $I_i, i = 1, \dots, N$ , of the sign language video into noisy latents  $x^i_t, i = 1, \dots, N$ , serving as starting points for the generation of anonymized video frames, following the method described by Meng et al. (2021). An anchor frame  $I_a$  is selected from the sequence  $I_i, i = 1, \dots, N$ . The corresponding latent feature  $x^a_t$ , along with the Holistically-Nested Edge, is processed through ControlNet to create the transformed anchor frame  $I'_a$ , which constrains the global consistency in general. Empirically, we find that selecting the anchor frame from the middle of the video, where both hands of the signer are visible, yields optimal results. For each frame  $I_i$ , the previously generated frame  $I'_{i-1}$  and the anchor frame  $I'_a$  provide cross-frame attention control during the generation of  $I'_i$ , as detailed in Section 3.2.1. A two-stage optical-flow-guided latent fusion, described in Section 3.2.2, is applied during the generation process. Finally, a specialized facial expression enhancement module, outlined in Section 3.3, is used to refine the results.

##### 3.2.1. Cross-Frame Attention Consistency

In the Stable Diffusion model, there are two kinds of attention mechanisms used in the U-Net. The cross-attention retrieves the information from the text embedding. The self-attention helps define the layout and style of the generated images. In order to achieve consistent generation across frames



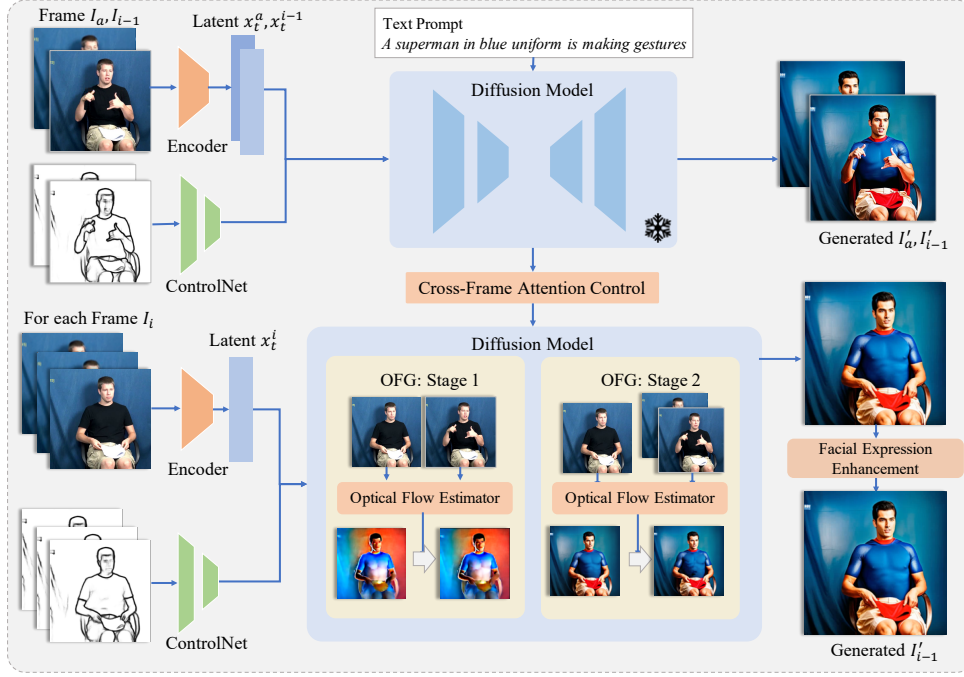


Figure 2: **Method Overview.** The original frames  $\{I_i\}$ ,  $i = 1, \dots, N$  in the sign language video are encoded and sampled as noisy latent features  $\{x_t^i\}$ ,  $i = 1, \dots, N$ . An anchor frame  $I_a$  and its Holistically-Nested Edge are used to generate the  $I'_a$  with ControlNet, which will constrain the global style consistency. For each frame  $I_i$ , the previous generated frame  $I'_{i-1}$  and the anchor-generated frame  $I'_a$  provide cross-frame attention control during the generation process of  $I'_i$ . A two-stage optical-flow-guided latent fusion is applied. A specialized facial expression enhancement module is used to update  $I'_i$  for the final result.

in the sign language video sequence, the self-attention layers are replaced with cross-frame attention layers. The self-attention layer of the U-Net used in Stable Diffusion is represented as follows:

$$Q = W^Q v_i, K = W^K v_i, V = W^V v_i, \quad (5)$$

where  $v_i$  is the latent features input to the self-attention layer when generating  $I'_i$ .  $W^Q$ ,  $W^K$ , and  $W^V$  are the weights for project  $v_i$  to the query, key, and value in the attention mechanism, respectively. The attention map  $SA$  is calculated as following:

$$SA(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (6)$$

where  $d$  is the dimension of  $K$ . To obtain consistent generation across frames, we replace  $K$  and  $V$  with  $K_{a,i-1}$  and  $V_{a,i-1}$ , which are the combination of keys and values when generating the selected anchor frame  $I_a$  and previous frame  $I_{i-1}$ . The cross-frame attention layer is represented as:

$$\begin{aligned} K_{a,i-1} &= W^K [v_a; v_{i-1}], & Q &= W^Q v_i \\ V_{a,i-1} &= W^V [v_a; v_{i-1}], \end{aligned} \quad (7)$$

where  $v_a$ ,  $v_{i-1}$  are the latent features obtained when generating frame  $I'_a$  and  $I'_{i-1}$ . The cross-attention map  $CA$  is calculated as:

$$CA(Q, K_{a,i-1}, V_{a,i-1}) = \text{Softmax}\left(\frac{QK_{a,i-1}^T}{\sqrt{d}}\right)V_{a,i-1} \quad (8)$$

The cross-frame attention mechanism is designed to foster consistency in image generation across frames by directing the current generation process to reference patches in both the generated anchor frame and the previous frame.

### 3.2.2. Optical-Flow-Guided Cross-Frame Latent Fusion

Following Yang et al. (2023), we use 2-stage latent fusion guided by optical flow: OFG stages 1 and 2.

- OFG stage 1: In the early stage of the diffusion backward process, the optical flow  $w_a^i$  and occlusion mask  $M_a^i$  are estimated from  $I_a$  to  $I_i$  to wrap and fuse the estimated latent of  $I'_a$  and  $I'_i$ . This latent wrap and fusion is performed when the denoising step  $t$  is large, to prevent distortion of results. At time step  $t$ , the predicted  $\hat{x}_0$  is updated by:

$$\hat{x}_0^i = M_a^i \hat{x}_0^i + (1 - M_a^i) w_a^i (\hat{x}_0^a), \quad (9)$$

where  $\hat{x}_0^i$  and  $\hat{x}_0^a$  are the predicted clear outputs for  $I'_i$  and  $I'_a$  at denoising time step  $t$ , from equation 4.

- OFG stage 2: At the second stage, the generated anchor frame  $I'_a$  and previous generated frame  $I'_{i-1}$  are used to further enhance consistency during the late stages of the diffusion backward process. The



optical flow and occlusion mask are also estimated. We obtain a reference image  $\bar{I}'_i$  by wrapping and fusing with the previous generated images:

$$\bar{I}'_i = M_a^i (M_{i-1}^i \hat{I}'_i + (1 - M_{i-1}^i) w_{i-1}^i (I'_{i-1})) + (1 - M_a^i) w_a^i I'_a, \quad (10)$$

After obtaining this reference-estimated image  $\bar{I}'_i$ , we can update the sampling process for generating  $I'_i$  using the following equation:

$$x_{t-1}^i = M_i x_{t-1}^i + (1 - M_i) \bar{x}_{t-1}^i, \quad (11)$$

where  $M_i = M_a^i \cap M_{i-1}^i$ , and  $\bar{x}_{t-1}^i$  is the sampled  $x_{t-1}^i$  from reference image  $\bar{I}'_i$ . We use the same strategy as the fidelity-oriented image encoding in Yang (2023) to encode  $\bar{I}'_i$  to avoid information loss when repeatedly encoding and decoding latents.

To maintain coherent color throughout the whole process, we also apply AdaIN (Huang and Belongie, 2017) to  $\hat{x}_0^i$  with  $\hat{x}_0^a$  at time step  $t$  during the late stage of the diffusion backward process. This mitigates the color drift problem with diffusion models.

### 3.3. Facial Expression Enhancement

Facial expressions convey important linguistic information in signed languages. However, current methods cannot transfer meaningful facial expressions; see the ablation study discussed in Section 4.6. ControlNet and Stable Diffusion usually fail to produce faces with the same expressions as the original signer. To address this issue, we propose an additional module to enhance the face generation based on an image-animation model. See Figure 3 for an overview of this module.

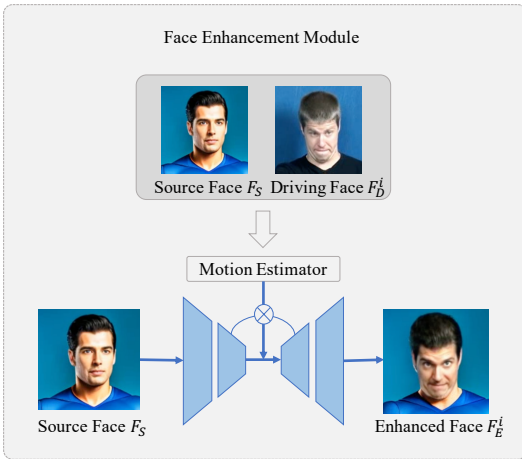


Figure 3: **Face Enhancement Module.** The motion estimator obtains dense motion and multi-resolution occlusion maps between the source face  $F_s$  and the driving face. The output along with a U-Net is applied to generate the enhanced face  $F_E^i$ .

When generating the first frame  $I'_1$ , we crop the result face and use it as the source face  $F_s$  for

the image animation module from Zhao and Zhang (2022). The facial images in the original videos are also cropped and aligned to formalize the driving face set  $[F_d^i], i = 1 \dots N$ . A motion estimation module will estimate the dense motion  $W_i$  and multi-resolution occlusion maps  $M_i$  between the source face  $F_s$  and the driving face set  $[F_d^i], i = 1 \dots N$ .

The obtained optical flow and occlusion maps are input to a U-Net to generate new face images that match the identity of the source face  $F_s$  but have the same facial expression as  $F_d^i$ . The input image  $F_s$  is processed through the encoder, and optical flow  $W_i$  is applied to wrap the feature map at each level. This adjusted feature map is then combined with the occlusion mask  $M_i^f$  that matches its resolution. Subsequently, it is merged into the decoder through a skip connection. The feature map is then input to the next upsampling layer. Finally, the enhanced face image  $F_E^i$  is produced at the last layer.

A face parser model (Yu et al., 2018) is applied on  $F_E^i$  to segment the face area and obtain a mask  $M_i^f$ . Then, the mask and enhanced face image are aligned with the face location in  $I'_i$ . Finally,  $I'_i$  is updated by the following equation:

$$I'_i = M_i^f F_E^i + (1 - M_i^f) I'_i. \quad (12)$$

## 4. Experiments and Results

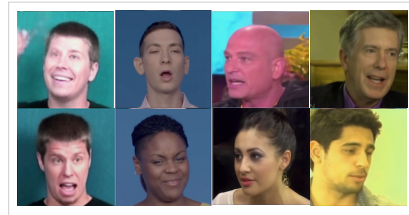


Figure 4: **Example Images** from the mixed dataset. We sampled more images from ASL videos for a balanced dataset.

### 4.1. Dataset

We implemented our method on video datasets distributed through the American Sign Language Linguistic Research Project (ASLLRP): <https://dai.cs.rutgers.edu/dai/s/dai> (Neidle et al., 2018, 2022b). Each test sample was limited to a maximum of 180 video frames. Example results are presented in Figure 5. We also produce a mixed dataset for fine-tuning the facial expression module, as illustrated in Section 4.3.

### 4.2. Models

Our experiments utilized Stable Diffusion models version 1.5 and other customized models. The ControlNet version 1.0 was employed, producing optimal results with HED as a conditional input. Optical flow estimation was performed using the model from Xu et al. (2022).

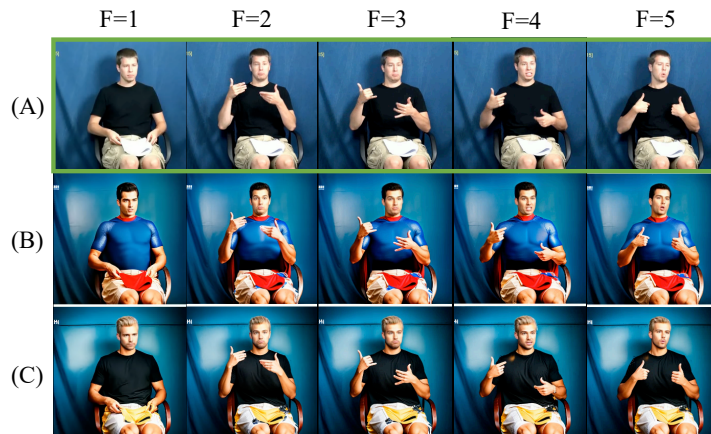


Figure 5: **Anonymization Result Examples.** Row (A) contains some frames from the original ASL video (taken from ASLLRP file Cory\_2013-6-27\_sc115, Utterance 22, meaning ‘If friends play Frisbee, I will join them.’). Rows (B) and (C) show anonymization from different prompts: (B) *a Superman in blue uniform is making gestures* (C) *a man in CG style, blond hair, is making gestures*.

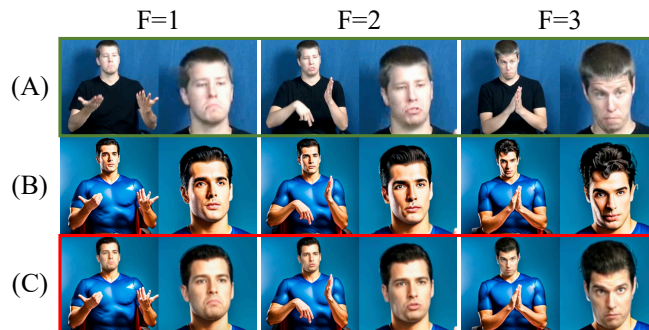


Figure 6: **Ablation Study of Facial Expression Enhancement.** The frames in Row (A) are taken from ASLLRP file Cory\_2013-6-27\_sc114, Utterance 102. Row (B) is the result without the facial enhancement module. Row (C) is the final result of our method.

### 4.3. Fine-tuning Facial Expression Model

State-of-the-art facial reenactment models are usually trained on large-scale speaking head datasets such as Voxceleb (Nagrani et al., 2017). The rich identity information contained in such datasets makes it possible to generalize on face images in the wild. However, the speaking head videos lack linguistically important facial expressions. In contrast, the face images cropped from ASL videos contain linguistic information, but lack diversity of identities, which impacts the model’s ability to generalize. To address this, we propose to mix these two datasets and apply a balance sampling strategy in training in order to maintain the model’s generalization ability and enable generation of facial expressions carrying linguistic meanings. Figure 4 shows example face images for this mixed dataset. We fine-tune the pre-trained model from Zhao and Zhang (2022) on this mixed dataset for 40 epochs.

### 4.4. Qualitative Evaluation

To our knowledge, this is the first instance of text-guided sign language anonymization capable of generating an unlimited array of diverse

anonymized videos. Methods like Cartoonized Anonymization (CA) (Tze et al., 2022b) cannot generate photorealistic results and rely on skeleton estimation for accurate anonymization. Methods that can generate photorealistic results, e.g., AnonySign (Saunders et al., 2021), SLA (Xia et al., 2022), and Neural Sign Reenactor (NSR) (Tze et al., 2022a), require accurate skeleton estimation or have very limited choices of anonymization identities.

Our initial results are encouraging. Our method can generate clear handshapes with high fidelity to the original signer’s handshapes and hand/arm movements. Most generated facial expressions are good; further refinements to fully preserve subtle linguistic expressions are underway. Effectiveness for complete disguise of identity, transmission of linguistic content, and production of natural-looking signing remains to be confirmed through user studies, to be carried out soon. In the very near future, we will also validate our results by processing our anonymized videos through our independent system for sign recognition from video (Zhou et al., 2024, under review), to confirm that the anonymized versions are correctly recognized

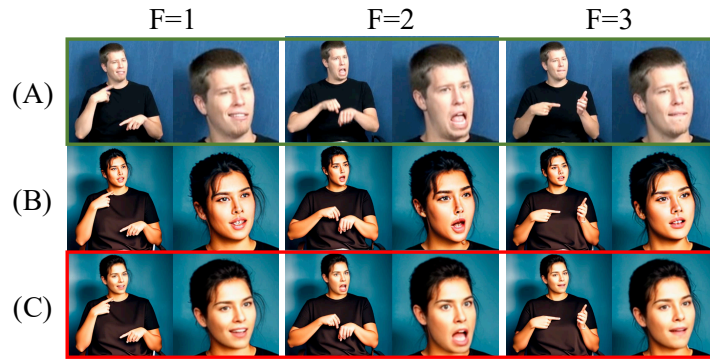


Figure 7: **Ablation Study of Facial Expression Enhancement.** The frames in (A) are taken from ASLLRP file Cory\_2013-6-27\_sc107, Utterance 14. Row (B) shows the result without the facial enhancement module. (C) shows the final result of our method.

as the originally produced sign. Figure 5 shows that our method can produce computer-generated signers with varying identities: Text prompts allow for varying anonymized versions of ASL videos. The results underscore the practical potential of our approach. Video examples can be seen at <https://github.com/Jeffery9707/DiffSLVA2>.

#### 4.5. Quantitative Evaluation

We use an identity classifier (Schroff et al., 2015; Cao et al., 2018) to check whether our method successfully changes the identity of the original signer. In particular, we calculate the cosine similarity between face embeddings of multiple images of the same signer and of anonymized signers. See Table 1. Cosine similarity close to 1 or 0 means the faces are from the same person or an unrelated person, respectively.

	Original	Anonymized
Signer A	0.7740	0.1273
Signer B	0.8917	0.0566
Signer C	0.8566	-0.0165

Table 1: Anonymization Analysis for the Face. Each column contains the cosine similarity between faces of the same signer and anonymized signers.

From the table, we can see that our anonymized face has a cosine similarity close to 0 with the original face. Therefore, our method has successfully anonymized the signers to a unrelated identity.

#### 4.6. Ablation Study

Our ablation study focused on the facial expression enhancement module. Results are shown in Figures 6 & 7. Using this module significantly improves preservation of linguistic meaning. (The examples shown include topic and wh-question marking.)

The Stable Diffusion model does not do well with accurate generation of varied facial expressions for ASL anonymization. Instead of producing diverse

expressions, the model tends to replicate a uniform expression across frames, resulting in loss of linguistic information. This limitation highlights the importance of applying facial expression enhancement module for ASL video anonymization.

## 5. Conclusion and Discussion

We introduce DiffSLVA, a novel approach using large-scale pre-trained diffusion models for text-guided ASL video anonymization. Our approach could be applied to various use cases. It could enable signers to share sensitive information while preserving privacy. It could enable anonymous peer review for ASL-based academic submissions, thereby ensuring unbiased academic review. It could bring neutrality to multimodal ASL tools, e.g., for anonymized definitions for ASL dictionaries. Furthermore, our approach could enhance neutrality in interpreting scenarios in digital communications, such as messaging, enabling maintenance of confidentiality in ASL communications. The implementation of DiffSLVA could also increase participation in video-based AI databases, enriching AI research with diverse ASL data.

This approach does not address the possibility that even anonymized signers could be recognized by those who know them very well, based on signing style. Furthermore, our current method has some limitations. It may encounter challenges in cases where the face is occluded by one or both hands or where there is blurring due to rapid movements in ASL videos. In addition, as is a known issue for Stable Diffusion Models, artifacts of various types sometimes appear in our anonymized videos. We aim to address these issues in our future work. We are also working on further refinements to improve the facial transformation module. However, overall, DiffSLVA shows substantial promise for anonymization applications, which could offer invaluable tools for the Deaf and Hard-of-Hearing communities.



## 6. Acknowledgments

We are grateful to the many, many people who have helped with the collection, linguistic annotation, and sharing of the ASL data upon which we have relied for this research. In particular, we are indebted to the many ASL signers who have contributed to our database; to Gregory Dimitriadis at the Rutgers Laboratory for Computer Science Research, the principal developer of SignStream®, our software for linguistic annotation of video data (<https://www.bu.edu/asllrp/SignStream/3/>); to Matt Huenerfauth and his team for data collection at RIT; to DawnSignPress for sharing video data; to the many who have helped with linguistic annotations (especially Carey Ballard and Indya Oliver); and to Augustine Opoku, for development and maintenance of our Web-based database system for providing access to the linguistically annotated video data (<https://dai.cs.rutgers.edu/dai/s/dai>). This work was supported in part by NSF grants #2235405, #2212302, #2212301, and #2212303, although any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## 7. Bibliographical References

- Robert W Arnold. 2009. *A proposal for a written system of American Sign Language*. Gallaudet University.
- Charlotte Baker-Shenk. 1985. The facial behavior of deaf signers: Evidence of a complex language. *American Annals of the Deaf*, 130(4):297–304.
- Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. *Demystifying MMD GANs*. *arXiv preprint arXiv:1801.01401*.
- Julian Bleicken, Thomas Hanke, Uta Salden, and Sven Wagner. 2016. *Using a language technology infrastructure for German in order to anonymize German Sign Language corpus data*. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3303–3306, Portorož, Slovenia. European Language Resources Association (ELRA).
- Danielle Bragg, Oscar Koller, Mary Bellard, Larian Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019a. *Sign language recognition, generation, and translation: An interdisciplinary perspective*. In *The 21st international ACM SIGACCESS conference on computers and accessibility*, pages 16–31.
- Danielle Bragg, Oscar Koller, Naomi Caselli, and William Thies. 2019b. *Exploring collection of sign language datasets: Privacy, participation, and model performance*. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14.
- Necati Cihan Camgöz, Ben Saunders, Guillaume Rochette, Marco Giovanelli, Giacomo Inches, Robin Nachtrab-Ribback, and Richard Bowden. 2021. *Content4all open research sign language translation datasets*. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE.
- Chen Cao, Qiming Hou, and Kun Zhou. 2014. *Displaced dynamic expression regression for real-time facial tracking and animation*. *ACM Transactions on graphics (TOG)*, 33(4):1–10.
- Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. *VGGFace2: A dataset for recognising faces across pose and age*. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74.
- Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. *Pix2video: Video editing using image diffusion*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217.
- Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. *Everybody dance now*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942.
- Geoffrey Restall Coulter. 1979. *American Sign Language typology*. Ph.D. thesis, University of California, San Diego.
- DawnSign Press. 2022. *DawnSign-Press (2022) About Us*. *DawnSignPress Website*.
- Eleni Efthimiou, Stavroula-Evita Fotinea, Theodore Goulas, and Panos Kakoulidis. 2015. *User friendly interfaces for sign retrieval and sign synthesis*. In *International Conference on Universal Access in Human-Computer Interaction*, pages 351–361. Springer.
- Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and François Lefebvre-Albaret. 2012. *The dicta-sign Wiki: Enabling web communication for the deaf*. In *International Conference on Computers for Handicapped Persons*, pages 205–212. Springer.



- Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. Sign language recognition, generation, and modelling: a research effort with applications in deaf communication. In *International Conference on Universal Access in Human-Computer Interaction*, pages 21–30. Springer.
- Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. [RMPE: Regional multi-person pose estimation](#). In *ICCV*.
- Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Šykora. 2017. [Example-based synthesis of stylized facial animations](#). *ACM Transactions on Graphics (TOG)*, 36(4):1–11.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. [Generative adversarial nets](#). *Advances in neural information processing systems*, 27.
- Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. 2020. [MarioNETte: Few-shot face reenactment preserving identity of unseen targets](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10893–10900.
- Alexis Heloir and Fabrizio Nunnari. 2016. [Toward an intuitive sign language animation authoring system for the deaf](#). *Universal Access in the Information Society*, 15(4):513–523.
- Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. [Prompt-to-prompt image editing with cross attention control](#). *arXiv preprint arXiv:2208.01626*.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. [GANs trained by a two time-scale update rule converge to a local nash equilibrium](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. [Denosing diffusion probabilistic models](#). *Advances in neural information processing systems*, 33:6840–6851.
- Xun Huang and Serge Belongie. 2017. [Arbitrary style transfer in real-time with adaptive instance normalization](#). In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510.
- Amy Isard. 2020. [Approaches to the anonymisation of sign language corpora](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 95–100, Marseille, France. European Language Resources Association (ELRA).
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. [Perceptual losses for real-time style transfer and super-resolution](#). In *European conference on computer vision*, pages 694–711. Springer.
- Hernisa Kacorri and Matt Huenerfauth. 2016. [Continuous profile models in ASL syntactic facial expression synthesis](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2084–2093.
- Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. [Text2video-zero: Text-to-image diffusion models are zero-shot video generators](#). *arXiv preprint arXiv:2303.13439*.
- Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 107–114.
- Sooyeon Lee, Abraham Glasser, Becca Dingman, Zhaoyang Xia, Dimitris Metaxas, Carol Neidle, and Matt Huenerfauth. 2021. [American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users](#). In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA. Association for Computing Machinery.
- Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. 2018. [CrowdPose: Efficient crowded scenes pose estimation and a new benchmark](#). *arXiv preprint arXiv:1812.00324*.
- Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2023. [Video-p2p: Video editing with cross-attention control](#). *arXiv preprint arXiv:2303.04761*.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. [Mediapipe: A framework for building perception pipelines](#). *arXiv preprint arXiv:1906.08172*.

- Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2021. [SDEdit: Guided image synthesis and editing with stochastic differential equations](#). *arXiv preprint arXiv:2108.01073*.
- Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. [Null-text inversion for editing real images using guided diffusion models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047.
- Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. [VoxCeleb: a large-scale speaker identification dataset](#). *arXiv preprint arXiv:1706.08612*.
- Carol Neidle, Judy Kegl, Benjamin Bahan, Dawn MacLaughlin, and Robert G Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT press.
- Carol Neidle, Augustine Opoku, Carey M. Ballard, Konstantinos M. Dafnis, Evgenia Chroni, and Dimitris Metaxas. 2022a. [Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 165–172, Marseille, France. European Language Resources Association (ELRA).
- Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. 2018. [New shared & interconnected ASL resources: SignStream® 3 software; DAI 2 for web access to linguistically annotated video corpora; and a sign bank](#). In *Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*, pages 147–154, Miyazaki, Japan. European Language Resources Association (ELRA).
- Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022b. [ASL Video Corpora & Sign Bank: Resources available through the American Sign Language Linguistic Research Project \(ASLLRP\)](#). *arXiv preprint arXiv:2201.07899*.
- Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. 2022. [On aliased resizing and surprising subtleties in gan evaluation](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11400–11410.
- Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. 2023. [FateZero: Fusing attentions for zero-shot text-based video editing](#). *arXiv preprint arXiv:2303.09535*.
- Shengju Qian, Kwan-Yee Lin, Wayne Wu, Yangxiaokang Liu, Quan Wang, Fumin Shen, Chen Qian, and Ran He. 2019. [Make a face: Towards arbitrary high fidelity face manipulation](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10042.
- Jian Ren, Menglei Chai, Sergey Tulyakov, Chen Fang, Xiaohui Shen, and Jianchao Yang. 2020. [Human motion transfer from poses in the wild](#). In *European Conference on Computer Vision*, pages 262–279.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. [High-resolution image synthesis with latent diffusion models](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. [AnonySIGN: Novel human appearance synthesis for sign language video anonymisation](#). *arXiv preprint arXiv:2107.10685*.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. [FaceNet: A unified embedding for face recognition and clustering](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [Laion-5b: An open large-scale dataset for training next generation image-text models](#). 35:25278–25294.
- Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. 2023. [Edit-a-video: Single video editing with object-aware consistency](#). *arXiv preprint arXiv:2303.07945*.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019a. [Animating arbitrary objects via deep motion transfer](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386.
- Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019b. [First order motion model for image animation](#). *Advances in Neural Information Processing Systems*, 32:7137–7147.

- Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. 2021. [Motion representations for articulated animation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662.
- Jenny L Singleton, Gabrielle Jones, and Shilpa Hanumantha. 2014. Toward ethical research practice with deaf participants. *Journal of Empirical Research on Human Research Ethics*, 9(3):59–66.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. [Denoising diffusion implicit models](#). *arXiv preprint arXiv:2010.02502*.
- Christina O Tze, Panagiotis P Filntisis, Athanasia-Lida Dimou, Anastasios Roussos, and Petros Maragos. 2022a. [Neural sign reenactor: Deep photorealistic sign language retargeting](#). *arXiv preprint arXiv:2209.01470*.
- Christina O Tze, Panagiotis P Filntisis, Anastasios Roussos, and Petros Maragos. 2022b. [Cartoonized anonymization of sign language videos](#). In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE.
- Clayton Valli and Ceil Lucas. 2000. *Linguistics of American Sign Language: An introduction*. Gallaudet University Press.
- Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. 2003. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018a. [Video-to-video synthesis](#). *arXiv preprint arXiv:1808.06601*.
- Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018b. [High-resolution image synthesis and semantic manipulation with conditional gans](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.
- Wen Wang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. 2023. [Zero-shot video editing using off-the-shelf image diffusion models](#). *arXiv preprint arXiv:2303.17599*.
- Olivia Wiles, A Koepke, and Andrew Zisserman. 2018. [X2Face: A network for controlling face generation using images, audio, and pose codes](#). In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–686.
- Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. [Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633.
- Zhaoyang Xia, Yuxiao Chen, Qilong Zhangli, Matt Huenerfauth, Carol Neidle, and Dimitris Metaxas. 2022. [Sign language video anonymization](#). In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 202–211, Marseille, France. European Language Resources Association (ELRA).
- Saining Xie and Zhuowen Tu. 2015. [Holistically-nested edge detection](#). In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403.
- Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. 2018. Pose Flow: Efficient online pose tracking. In *BMVC*.
- Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, and Dacheng Tao. 2022. [GMFlow: Learning optical flow via global matching](#). In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130.
- Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. 2023. [Rerender a video: Zero-shot text-guided video-to-video translation](#). In *ACM SIGGRAPH Asia Conference Proceedings*.
- Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. [Bisenet: Bilateral segmentation network for real-time semantic segmentation](#). In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. [Few-shot adversarial learning of realistic neural talking head models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9459–9468.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. [Adding conditional control to text-to-image diffusion models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Jian Zhao and Hui Zhang. 2022. [Thin-plate spline motion model for image animation](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666.

Yang Zhou, Zhaoyang Xia, Yuxiao Chen, Carol Neidle, and Dimitris N. Metaxas. 2024. A Multimodal Spatio-temporal GCN Model with Enhancements for Isolated Sign Recognition. In *Proceedings of the LREC2024 11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*.



# A Multimodal Spatio-Temporal GCN Model with Enhancements for Isolated Sign Recognition

Yang Zhou<sup>1</sup>, Zhaoyang Xia<sup>1</sup>, Yuxiao Chen<sup>1</sup>, Carol Neidle<sup>2</sup>, Dimitris N. Metaxas<sup>1</sup>

<sup>1</sup> Rutgers University, <sup>2</sup> Boston University

<sup>1</sup> 110 Frelinghuysen Road, Piscataway, NJ 08854,

<sup>2</sup> Boston University Linguistics, 621 Commonwealth Ave., Boston, MA 02215

{eta.yang, zx149}@rutgers.edu, yc984@cs.rutgers.edu, carol@bu.edu, dnm@cs.rutgers.edu

## Abstract

We propose a multimodal network using skeletons and handshapes as input to recognize individual signs and detect their boundaries in American Sign Language (ASL) videos. Our method integrates a spatio-temporal Graph Convolutional Network (GCN) architecture to estimate human skeleton keypoints; it uses a late-fusion approach for both forward and backward processing of video streams. Our (core) method is designed for the extraction—and analysis of features from—ASL videos, to enhance accuracy and efficiency of recognition of individual signs. A Gating module based on per-channel multi-layer convolutions is employed to evaluate significant frames for recognition of isolated signs. Additionally, an auxiliary multimodal branch network, integrated with a transformer, is designed to estimate the linguistic start and end frames of an isolated sign within a video clip. We evaluated performance of our approach on multiple datasets that include isolated, citation-form signs and signs pre-segmented from continuous signing based on linguistic annotations of start and end points of signs within sentences. We have achieved very promising results when using both types of sign videos combined for training, with overall sign recognition accuracy of 80.8% Top-1 and 95.2% Top-5 for citation-form signs, and 80.4% Top-1 and 93.0% Top-5 for signs pre-segmented from continuous signing.

**Keywords:** ASL, GCN, Gating module, Temporal action localization

## 1. Introduction

In the US, it is estimated that 28 million people are Deaf or hard of hearing (Lin et al., 2011), and that about 500,000 use American Sign Language (ASL) as their primary language (Mitchell et al., 2006). ASL is also the 3rd most studied non-native language (Looney and Lusin, 2021). Signed languages are full-fledged natural languages, with information expressed in the visual-gestural modality by movements of the arms, hands, head, and upper body, and by facial expressions. They generally lack a standardized written form.

Computer-aided sign language analytics and sign recognition from video have many potential applications, which include resources to provide/enhance access to digital materials for signers, and tools for sign language learners (including hearing parents of deaf children) and interpreters, for ASL-to-English translation, and for improved sign language research. Research in this area is challenging, however, in part because of the complexity and variability of sign production and the fact that information expressed across the relevant channels may differ in spatio-temporal scale. For example, grammatical information conveyed non-manually by facial expressions and head gestures may extend over phrasal domains, i.e., it may occur over a scope that includes more than one sign. In this paper, we focus on the recognition of individual signs—both isolated, citation-form signs

and signs pre-segmented from continuous signing. This is a critical step towards recognition of signs directly from sentences. Sign production in continuous signing differs somewhat from production of citation-form signs (Neidle, 2023), so it is particularly significant that we are able to achieve a high degree of success also for recognition of pre-segmented signs trained on the combined dataset.

One major challenge is the existence of both inter- and intra-signer variations in sign production. Another significant challenge results from the fact that different classes of signs (e.g., lexical signs, fingerspelled signs, and classifiers) have significantly different internal structures. Addressing these challenges requires extensive video datasets with diverse signers and consistent gloss labeling of signs, to train computational models effectively. We utilize multiple datasets shared on the Web by the American Sign Language Linguistic Research Project (ASLLRP) (Neidle et al., 2022b)—specifically, their collections of **isolated, citation-form signs** (ASLLVD (Neidle and Metaxas, 2023b), DSP (Neidle and Metaxas, 2023c), and RIT (Neidle and Metaxas, 2023e)), and of **signs pre-segmented from continuous signing** based on linguistic annotations that include information about the linguistic start and end points of these signs within sentences (ASLLRP Sentences (Neidle and Metaxas, 2023a) and DSP Sentences (Neidle and Metaxas, 2023d))—as well

isolated sign data from WLASL (Li et al., 2020), with annotations provided by ASLLRP to ensure consistent labeling (Neidle et al., 2022a; Neidle and Ballard, 2022). Taken together, this collection includes 21,083 videos with over 2,000 distinct signs from 119 signers, with consistent gloss labeling and a focus on lexical signs. This collection, which will be referred to in this paper as the “ASLLRP Individual Sign Collection,” forms the basis for our experiments to advance sign recognition using deep learning techniques.

Prior to the advent of deep learning methods, traditional machine learning methods such as Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs) were employed to capture the spatio-temporal aspects of sign language (Lafferty et al., 2001; Grobel and Assan, 1997; Dilsizian et al., 2014). Recent advances in deep learning, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), have opened new avenues towards the automated recognition of signs from large vocabularies without the manual identification of features in the video. However, several challenges remain. For example: (1) Many of the available video resources have poor spatio-temporal resolution; (2) There are many different types of signs, with different internal composition, and some types, such as classifier constructions (which incorporate some degree of iconicity) do not constitute a fixed vocabulary; (3) The size of the data is relatively small, compared to spoken-language datasets; and (4) There is no 1-1 correspondence between ASL signs and English words, and no agreed-upon convention for providing English-based gloss labels to uniquely identify ASL signs. In this paper, we present results for recognition of individual ASL lexical signs, using the largest-to-date dataset that includes both isolated signs and signs pre-segmented from continuous signing. For more precise recognition, we have also developed a new approach to detect the beginning and end of an isolated, citation-form sign within a video clip.

## 2. Overview of our Approach

To achieve accurate sign recognition from video, we propose a deep learning approach based on skeletons. This method involves detecting start and end frames of the signs, and it leverages parameters from the skeleton data. Using a bidirectional learning framework within a Graph Convolutional Network (GCN) architecture, our method achieves notable accuracy on the ASLLRP Individual Sign Collection and WLASL data.

To improve sign recognition accuracy for the set of isolated signs, a Gating module designed to

evaluate temporal weights has been embedded to enable the network to focus on the significant frames in the video clips, while avoiding frames that contain blurring or other artifacts often present in videos. To further enhance the feature extraction model, we designed an auxiliary multi-modal branch network for temporal action localization based on an encoder and transformers. With training based on linguistic annotations of start and end frames in the ASLLVD and DSP isolated sign datasets, the auxiliary branch utilizes spatio-temporal features extracted by the GCN and the encoded handshape information, to detect the start and end points of isolated signs. The resulting improvements in sign recognition accuracy are shown in Section 5.3.3.

## 3. Related Work

Before the advent of deep learning techniques, sign language recognition research relied primarily on handcrafted features, such as the positioning and movement of hands relative to specific body parts (Tornay et al., 2020; Cooper et al., 2012; Badhe and Kulkarni, 2015; Xiaohan Nie et al., 2015), combined with standard classifiers like Support Vector Machines (SVMs), k-Nearest Neighbors (kNNs), Conditional Random Fields (CRFs), and Hidden Markov Models (HMMs) (Memiş and Albayrak, 2013; Dardas and Georganas, 2011; Yang, 2010; Metaxas et al., 2018; Tornay et al., 2020). However, these handcrafted features and underlying Gaussian distribution assumptions limited the systems’ capabilities for generalization and scalability. Recently, deep neural network methods have made breakthroughs in computer vision tasks, such as action and gesture recognition; these methods have also been applied to sign language recognition, a more difficult problem given the complexity of linguistic structure (Rastgoo et al., 2021; Jiang et al., 2021). Some recent research has used transfer learning methods for isolated sign recognition, since available sign language datasets have vocabularies that are small compared to those of general-purpose human motion databases like Kinetics400 (Carreira and Zisserman, 2017). Such approaches are discussed by Sandoval-Castañeda et al. (2023), who attained best results using a visual transformer pretrained first on human action videos in Kinetics400, and then on OpenASL (Shi et al., 2022) videos (following Wei et al. (2022)). They fine-tuned on the WLASL (Li et al., 2020) dataset—with modified glossing (as in Dafnis et al., 2022b; Neidle et al., 2022a; Neidle and Ballard, 2022). They also leveraged phonological features extracted from ASL-LEX 2.0 (Sevcikova Sehyr et al., 2021), to “better characterize video models and pre-training tasks.” See further

discussion in Section 5.4.

### 3.1. RGB-based Approaches

In sign recognition, RGB-based approaches have undergone a significant evolution with the rise of deep learning. Initially, these methods focused on extracting spatial features from RGB frames using traditional image processing techniques. The introduction of Convolutional Neural Networks (CNNs) marked a significant advance, allowing for more efficient and nuanced extraction of spatial features directly from RGB data.

Pioneering work by Krizhevsky et al. (2012) and Simonyan and Zisserman (2014) showcased the effectiveness of CNNs in automated image feature extraction (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), laying the groundwork for applying these networks to sign language recognition. These CNN models are adept at analyzing shapes, movements, and orientations of hands and body parts, critical for sign recognition. However, the challenge in sign recognition extends beyond spatial to temporal feature extraction. This led to the integration of CNNs with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, known for their ability to capture temporal dynamics in sequences, as described by Hochreiter and Schmidhuber (1997). Further advances were achieved with 3D Convolutional Neural Networks (3D-CNNs), which, as explored by Ji et al. (2013), extract spatio-temporal features from video sequences, offering a more holistic approach to gesture recognition. More recent studies have investigated use of attention mechanisms, particularly in Transformer models (Vaswani et al., 2017), for sign recognition. These mechanisms focus on specific segments of video frames, enhancing recognition accuracy by highlighting critical sign language features.

Despite these technological advances, RGB-based methods still face challenges, in part because of sensitivity to lighting conditions, foreground-background complexities, and possible lack of focus on the important parts of the human body. This translates to an increased need for training data, which are unavailable in real-world settings. Our model-based approach aims to overcome these limitations, enhancing the robustness and applicability of sign language recognition systems in various real-world settings.

### 3.2. Skeleton-based Approaches

Skeleton-based approaches for action and sign language recognition have significantly evolved, focusing on extracting and analyzing body keypoints or skeleton graphs. Facilitated by advanced human pose estimation technologies, this

methodology prioritizes essential movement features while excluding irrelevant background noise. Initial research utilized Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture temporal aspects of actions (Soo Kim and Reiter, 2017; Liu et al., 2017). However, these models struggled with encoding spatial and temporal interactions between keypoints.

Addressing these limitations, Yan et al. (2018) introduced the Spatial Temporal Graph Convolutional Network (ST-GCN), showcasing the potential of Graph Convolutional Networks (GCNs) in learning skeleton dynamics. Despite this innovation, ST-GCNs, focusing on direct joint connections, overlooked critical indirect keypoint interactions, which are essential for comprehensive sign recognition. Efforts to surmount this challenge included Li et al.'s (2019a) exploration of latent connections and Shi et al.'s (2019b; 2020) multi-stream approaches that enhanced action recognition by integrating keypoints, bones, and their motion. Additionally, de Amorim et al. (2019) adapted the ST-GCN framework for sign recognition, achieving approximately 60% accuracy in recognizing a limited vocabulary of signs.

Further advances are exemplified by Jiang et al. (2021), which implemented a pose-based GCN with additional modalities like RGB frames and optical flow, resulting in significant progress in isolated sign recognition. Dafnis et al. (2022a) extended these approaches by incorporating forward and backward data streams with keypoints and bones acceleration, significantly improving recognition accuracy on the WLASL dataset.

## 4. Methodology

The human body can be represented as a graph with nodes consisting of the face, upper body, arms, and hands. For sign recognition, all these parts are important and need to be used. Therefore, our approach extracts this information from video based on the following three components: (1) a spatio-temporal Graph Convolutional Network (GCN) architecture, for detailed modeling of skeleton keypoints from a signer's video; (2) a late-ensemble technique to synergistically combine, in the GCN, the forward and backward video streams, for improved sign recognition; and (3) an Encoder and Transformer-based approach, for precise temporal motion localization of the beginning and end frames of a sign.

### 4.1. Spatio-temporal Graph Convolutional Network

Our goal is to capture and analyze the complex spatio-temporal movement dynamics of the arms



and hands during signing. To achieve this, our method first extracts keypoints and bones from the torso, arms, and hands using Alphapose, as developed by Fang et al. (2017). This method is capable of estimating 136 keypoints for the entire body from single RGB images. Using this model, we constructed a skeletal graph consisting of 27 nodes. These keypoints and respective bones are integrated within a GCN using spatio-temporal graph convolutions. Our model’s spatial convolutions are computed based on the spatial partitioning strategy described in the ST-GCN framework by Yan et al. (2018). The integration of spatio-temporal graph convolutions enables our model not only to capture the spatial relationships between keypoints and bones, but also to estimate their temporal evolution over time. This dual capability showcases the unique advantage of the ST-GCN framework in capturing both spatial intricacies and temporal variations. The spatial formulation of our GCN model is delineated as follows:

$$x_{\text{out}} = \Lambda^{-\frac{1}{2}}(I + A)\Lambda^{-\frac{1}{2}}x_{\text{in}}W, \quad (1)$$

where  $x_{\text{in}}$  in the GCN input consists of keypoints, bones, and other related information, while  $x_{\text{out}}$  denotes the output feature matrix derived from the graph convolution process. Matrix  $A$  models the intra-body connections (bones), while the identity matrix  $I$  models self-connections (keypoints).  $\Lambda$  is a diagonal matrix derived from  $(I + A)$ , and  $W$  is the ST-GCN weight matrix (2018). For purposes of our proposed application, the spatial graph convolutions are modeled using 2D convolution operations; the result,  $x_{\text{in}}W$ , is then multiplied by the normalized term  $\Lambda^{-\frac{1}{2}}(I + A)\Lambda^{-\frac{1}{2}}$  to compute  $x_{\text{out}}$

The right of Figure 1 shows the ST-GCN network architecture. Notably, a Gating module is appended to the end of the network, specifically focusing on important frames in isolated sign videos. The middle of Figure 1 illustrates the architecture of each of the GCN Blocks. It is composed of a Decoupled Spatial GC, STC Attention, a Temporal GC, and a series of Batch Normalization (BN) layers along with ReLU activation functions. The entire GCN Block includes a tail concatenation in the form of a residual structure to preserve low-level feature information. Drop Graphs are used in certain locations to prevent overfitting. The left part of Figure 1 provides details of the STC Attention Block, which consists of three attention modules: Spatial Attention, Temporal Attention, and Channel Attention, each with a tail concatenation to model the residual structure.

The Gating module in our approach is designed to identify and remove frames that are not useful for recognizing the sign, such as those with blurring or extraneous movements. We achieve this by

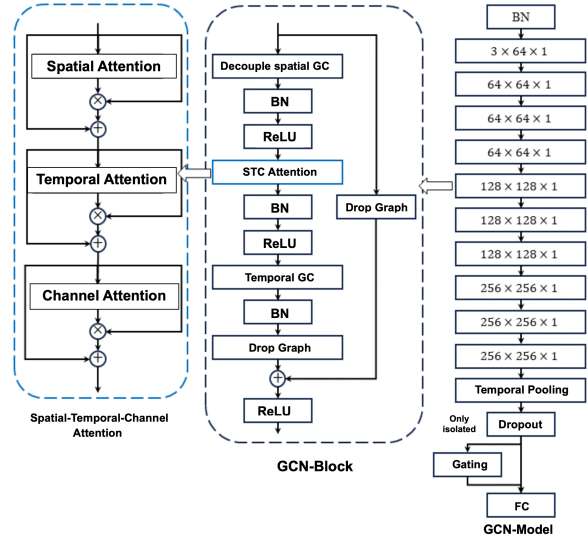


Figure 1: The ST-GCN Network Architecture

designing a multilayer convolution-based temporal attention module, to identify and remove those non-informative frames, as shown at the top of Figure 2. In this module, the skeleton feature dimension computed from the previous layers is reduced using a 3-layer stack of convolutions; a sequence of weights related to the temporal dimension is obtained by a temperature softmax layer (Hinton et al., 2015). The skeleton features computed from the previous layers are then multiplied with the output of the softmax layer in the Gating Block. Using this Gating Block, the network focuses, in the case of isolated signs, on those frames that carry valid information for sign recognition.

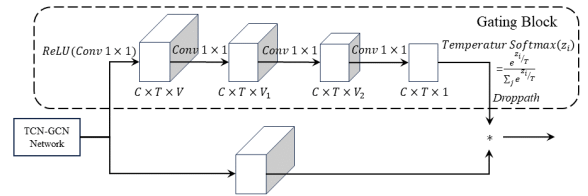


Figure 2: Gating Module Architecture

## 4.2. Bidirectional Stream GCN

Drawing inspiration from the multi-stream approach used in Shi et al. 2020, our methodology incorporates both forward and backward directions of video frame sequences for two types of data inputs: the location coordinates of the skeleton keypoints, and the bone vectors. To represent the bone vectors in our graph, we designate the nose as the root keypoint. Subsequent bone vectors are computed by tracing the connections between consecutive skeletal keypoints, starting from this root. As shown in Figure 3, the temporal data from the skeleton are processed with respect to two types of input: joints and bones; these are



then input into the forward stream. Subsequently, the temporal dimension is reversed and input into the backward stream. Then an ensemble from the predictions of the four models gives rise to a final prediction for the sign, as shown in Figure 3.

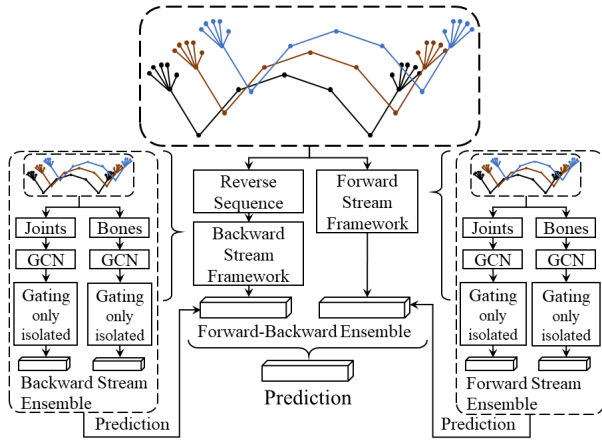


Figure 3: Bidirectional Stream GCN Architecture

#### 4.2.1. Score Fusion

As mentioned previously, our proposed framework uses two types of information streams, specifically joints and bones. We use their forward and backward directions to arrive at an improved consolidated prediction. We first integrate the prediction scores from these streams within each direction by using the softmax scores from each stream, as described by Shi et al. (2019a; 2019b; 2020); Cai et al. (2021); and Dafnis et al. (2022a), to calculate an optimized weighted sum of the scores pertinent to each direction. This process is then replicated for the fusion of prediction softmax scores from both directions; an optimized weighted summation is computed to predict the sign labels.

#### 4.2.2. Temporal Action Localization

To locate the start and end frames of isolated signs and thereby improve sign recognition, we design an auxiliary multimodal branch network. We train, using, in the loss function, linguistic annotations (which include the start and end frames of signs, and the handshapes in those frames), to learn to identify the start and end frames of a given isolated sign. As shown in Figure 4, the GCN network architecture is used to extract spatio-temporal features. Additionally, up to four types of handshapes for each sign video—Dominant start handshape, Dominant end handshape (and, for 2-handed signs, also Non-dominant start handshape and Non-dominant end handshape)—and the video are input into the network via a custom encoder. These are then processed through a transformer layer to improve the temporal positional dependence and interpretability of the hand-

shapes. The extracted features are concatenated with the features extracted by the GCN using a Temperature Softmax to predict the start and end frames of the isolated sign.

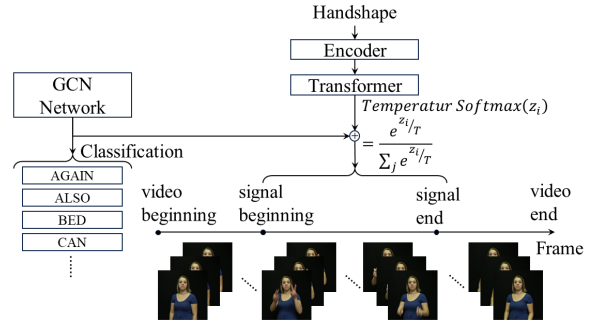


Figure 4: Auxiliary Multimodal Branch Architecture for Action Localization

## 5. Experiments

### 5.1. Data Preprocessing

Following the dataset partitioning strategy outlined in Dafnis et al. 2022b and Li et al. 2020, we divided the dataset into training, validation, and testing subsets. The division was carried out in a ratio of approximately 4:1:1 for each sign category; hence we further restrict these datasets to signs with at least 6 examples. For assessing the efficacy of sign recognition, we employed an evaluation metric based on the mean Top-K accuracy scores, where K is set to 1 and 5, applied across all instances of the signs.

We have used different combinations of the datasets for different tasks.

- To recognize isolated and pre-segmented sign videos, we combined video clips from all six datasets as follows: the **isolated** sign collections (WLASL (19,666 video clips), RIT (12,197 video clips), ASLLVD (9,746 video clips), DSP (2,935 video clips)); and the **pre-segmented** sign collections: ASLLRP (17,222 video clips) and DSP Sentences (hereafter referred to as DSP\_S, 3,136 video clips); totaling 64,902 video clips. After imposing a requirement of at least 6 available example video clips per sign, we arrived at a total of 56,681 distinct video clips corresponding to 2,377 distinct signs.
- To recognize isolated sign videos, the four isolated sign datasets just listed were used, with a total of 44,544 video clips. With the same restriction on example count, this yielded 41,597 distinct video clips corresponding to 2,295 distinct signs. We use the whole video clip, without estimating the beginning and the end frames of the sign.

- To train for recognition of the start and end frames of isolated signs, we merged the two isolated datasets for which we had ground truth annotations for the start and end frames of signs—ASLLVD and DSP—with a total of 12,681 distinct video clips corresponding to 748 distinct signs.

The process of graph construction begins with the normalization of keypoint coordinates within the range of  $[-1, 1]$ . We then apply a variety of data augmentation techniques, including random sampling, mirroring, rotation, scaling, and shifting. Considering the varying lengths of the videos, we standardize all videos to a uniform length of 200 frames. For videos exceeding this frame count, only the initial 200 frames are used. This truncation does not result in any significant loss of information because of the nature and length of the signs in our datasets. Conversely, for videos shorter than 200 frames, we pad zeros to the end of the temporal dimension to fill up to 200 frames.

## 5.2. Training Details

We employ Pytorch version 1.7.0 alongside a NVIDIA Quadro RTX8000 graphics card for all computational operations. The Graph Convolutional Network (GCN) models, encompassing both forward and backward streams, are trained under specific parameter settings. The training uses the Cross-Entropy loss function, with a finely-tuned weight decay parameter set to  $1 \times 10^{-4}$ . For optimization, Stochastic Gradient Descent (SGD) with Nesterov Momentum was the chosen method, where the momentum is maintained at 0.9. We initiated the learning rate at 0.1, reducing it by a factor of 10 at the 100th and 150th epoch milestones, culminating the training at 200 epochs.

With respect to batch processing, the batch size is uniformly set at 64 across both the training and testing stages. Each training iteration involves the random selection of 64 videos as inputs, ensuring a varied and comprehensive exposure of the dataset in each epoch. This strategy is pivotal in incorporating every video in the dataset into the training process, thus enhancing the robustness and diversity of the model training.

## 5.3. Results

### 5.3.1. GCN Performance

The sign recognition accuracy achieved using the combination of methods described in this paper is presented in Tables 1, 2, and 3.

### 5.3.2. Improvements in Performance Resulting from Use of Gating & Fusion

The score fusion of the forward and backward streams enhances overall sign recognition, as does the use of Gating for isolated sign video clips.

	WLASL	ASLLVD	RIT	DSP	Comb.
<i>Top-1</i>	79.59%	85.53%	75.98%	80.73%	<b>79.98%</b>
<i>Top-5</i>	95.32%	96.57%	93.22%	95.70%	<b>95.04%</b>

Table 1: Recognition accuracy for isolated signs trained on the combined isolated sign collections

	WLASL	ASLLVD	RIT	DSP	Comb.
<i>Top-1</i>	81.32%	86.70%	75.31%	79.97%	<b>80.76%</b>
<i>Top-5</i>	95.41%	96.95%	93.38%	95.28%	<b>95.18%</b>

Table 2: Recognition of isolated signs trained on the combined isolated & pre-segmented datasets

	ASLLRP	DSP_S	Comb.
<i>Top-1</i>	81.58%	73.86%	<b>80.39%</b>
<i>Top-5</i>	93.39%	90.62%	<b>92.96%</b>

Table 3: Recognition of pre-segmented signs trained on the combined isolated & pre-segmented datasets

This is shown in Table 4. The Bidirectional model’s Top-1 and Top-5 performance using forward and backward streams of joints and bones is presented in that table. The first four columns show recognition of isolated signs—based on training on the combined isolated sign collections—with and without Gating. The last two columns show results for recognition of signs from (and trained on) the combined isolated and pre-segmented datasets. It should be noted that the Gating module is not needed for our pre-segmented sign videos, since the start and end frames of these videos had been determined based on linguistic annotations of the start and end points of these signs.

### 5.3.3. Temporal Action Localization

In this section, we report (1) the accuracy of identification of the start and end frames of signs in isolated video clips, and then (2) the resulting improvement in sign recognition accuracy.

#### 1. Accuracy of Temporal Action Localization

To validate the accuracy of detection of start and end frames, we use the ASLLVD and DSP datasets—for which we have linguistic annotations of the start and end frames for signs. Table 5 presents the Mean Absolute Deviation (MAD), computed separately for the start and end frames as follows:

$$\text{MAD}_{\text{start}} = \frac{1}{N} \sum_{i=1}^N |p_{s_i} - g_{s_i}| \quad (2)$$

$$\text{MAD}_{\text{end}} = \frac{1}{N} \sum_{i=1}^N |p_{e_i} - g_{e_i}| \quad (3)$$

where,  $p_{s_i}$  and  $p_{e_i}$  are the predicted start and end frames for the  $i$ -th segment, while  $g_{s_i}$  and  $g_{e_i}$  are

	Isolated (no Gating)		Isolated (with Gating)		Isolated and Pre-segmented	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Forward stream of joints	74.04%	93.06%	74.82%	93.26%	75.59%	92.15%
Forward stream of bones	74.17%	92.63%	75.33%	93.12%	75.07%	92.52%
Backward stream of joints	73.36%	91.82%	73.96%	91.81%	74.02%	91.40%
Backward stream of bones	72.52%	92.44%	75.09%	92.71%	75.49%	92.28%
Fusion	79.24%	94.89%	79.98%	95.04%	80.61%	94.96%

Table 4: Recognition performance for forward and backward streams, where the isolated signs shown in the first 4 columns had been trained on the combined isolated sign data, and the combined isolated and pre-segmented signs in the final 2 columns had been trained on that total dataset

the annotated start and end frames for the  $i$ -th examples, and  $N$  is the total number of examples.

This is a measure of the deviation between the annotations and predictions for start and end frames of signs in videos with a frame rate of about 30 fps. However, it should be noted that in some cases, there is minimal difference in the images of the annotated and predicted frames; and in some other cases, the prediction may actually be more accurate than the annotation.

	start frame	end frame
ASLLVD	3.03	3.00
DSP	3.93	5.33
<b>Comb.</b>	3.24	3.56

Table 5: Mean Absolute Deviation between annotated and predicted start and end frames

## 2. Resulting Improvement in Sign Recognition

When our auxiliary multimodal branch network is used to segment signs in our isolated sign datasets, this results in some improvement in sign recognition rates. All video clips were subjected to segmentation processing prior to being input into the GCN model. Table 6 presents the recognition results for the isolated sign datasets, trained on the combined isolated sign datasets, by the GCN model WITH (row [2]) and WITHOUT (row [1]) prior segmentation.

	WLASL	ASLLVD	RIT	DSP	<b>Comb.</b>
[1] <i>Top-1</i>	79.41%	85.35%	75.72%	80.62%	<b>79.78%</b>
<i>Top-5</i>	95.15%	96.53%	93.11%	95.58%	<b>94.92%</b>
[2] <i>Top-1</i>	79.59%	85.53%	75.98%	80.73%	<b>79.98%</b>
<i>Top-5</i>	95.32%	96.57%	93.22%	95.70%	<b>95.04%</b>

Table 6: Sign recognition accuracy from isolated sign video clips: rows in [1] WITHOUT – and rows in [2] WITH – prior segmentation based on detected sign start and end frames

Although sign segmentation results directly in only a very slight improvement, there are additional ways in which we plan to leverage the ability to

identify the start and end frames of lexical signs, specifically with respect to explicit detection of handshapes. As demonstrated by Dilsizian et al. (2014), e.g., it is possible to exploit the linguistic dependencies that hold between start and end handshapes and between the handshapes on the two hands of lexical signs, to improve handshape recognition, which is an important component of sign recognition. They showed that incorporation of statistical information about such handshape dependencies, which can be derived from our annotated corpora, results in significant improvements in isolated sign recognition for lexical signs. This is planned for future research.

## 5.4. Comparisons of Overall Isolated Sign Recognition Accuracy

Table 7 compares the accuracy of our proposed model against state-of-the-art methods for recognition of signs from the WLASL dataset (Li et al., 2020). The overview at the top is taken from Xiao et al. (2023), Table 2 "Recognition performance comparison for different learning methods in WLASL dataset;" it shows results from [1] (Vinyals et al., 2016); [2] (Snell et al., 2017); [3] (Sung et al., 2018); [4] (Ravi and Larochelle, 2016); [5] (Mishra et al., 2017); [6] (Finn et al., 2017); [7] (Cai et al., 2018); [8] (Gidaris and Komodakis, 2018); [9] (Gordon et al., 2018); [10] (Qiao et al., 2018); [11] (Gidaris and Komodakis, 2019); [12] (Garcia and Bruna, 2017); [13] (Li et al., 2019b); [14] (Liu et al., 2018); and their own [15] (Xiao et al.). These studies used the WLASL dataset, which contains 21,083 video clips with about about 2,000 ASL signs.

As shown at the bottom of the table, our model secured the highest recognition rates for both Top-1 and Top-5. However, it should be noted that Dafnis et al. (2022b) and our own research used a partial but substantial subset of the WLASL data, consisting of 19,672 video examples, reglossed to ensure consistency of labeling (both internal to the WLASL dataset and across our other datasets (Neidle et al., 2022a; Neidle and Ballard, 2022)).

**OVERVIEW from Xiao et al. (2023)**

Method	Top-1	Top-5
<i>Metric-based</i>		
Matching Nets [1]	41.22%	50.26%
Prototypical Nets [2]	47.61%	65.13%
Relation Net [3]	45.26%	63.21%
<i>Meta-based</i>		
MetaLSTM [4]	41.56%	60.38%
SNAIL [5]	42.18%	53.77%
MAML [6]	46.21%	59.15%
MMNet [7]	52.13%	65.06%
Dynamic-Net [8]	54.21%	70.21%
<i>Generation-based</i>		
VERSA [9]	49.11%	61.19%
Param Predict [10]	55.36%	73.28%
wDAE [11]	55.05%	70.12%
<i>Graph-based</i>		
GNN [12]	52.02%	63.89%
CovaMNet [13]	51.18%	66.39%
TPN [14]	52.15%	65.22%
SL-GCN [15]	56.15%	73.26%

**COMPARE WITH**

Dafnis et al. 2022b	77.43%	94.54%
<b>Ours</b>	79.59%	95.32%

Table 7: Performance on the WLASL dataset (which contains isolated signs)

Sandoval-Castañeda et al. (2023) also used this subset of the WLASL dataset, with the same revised glosses. Using a very different approach (summarized in Section 3), they obtained similar results, with 79.02 % Top-1 recognition accuracy; Top-5 accuracy was not reported.

Table 8 compares performance of our model, with training on our isolated sign collection, and that of Dafnis et al. (2022b) on the same combined WLASL and ASLLVD dataset. We attained an improvement of 2.86% in Top-1 accuracy.

Combined	WLASL & ASLLVD	
	Top-1	Top-5
Dafnis et al. 2022b	78.70%	94.79%
<b>Ours</b>	81.56%	95.73%

Table 8: Performance on the same combined WLASL &amp; ASLLVD datasets

**6. Conclusions**

We introduce here a comprehensive framework for recognition of individual ASL signs. Although most prior related research has focused on isolated, citation-form signs, we successfully extend our recognition to include signs pre-segmented from continuous signing. Our method relies on

spatio-temporal GCNs, enhanced by bidirectional stream processing, and, for isolated signs, introduction of a Gating module and an auxiliary multimodal branch for temporal action localization. Our methodology addresses many of the inherent challenges of sign language recognition.

The application of our framework to an extensive collection of different datasets results in a high degree of recognition accuracy. For present purposes, we have used only a limited set of information from facial expressions (i.e., skeleton key-points), to establish a baseline. In future work we will explore adding more complete information from facial expressions, as this has been shown to improve sign recognition accuracy (von Agris et al., 2008).

We achieve state-of-the-art performance across various metrics, with overall sign recognition accuracy of 80.8% Top-1 and 95.2% Top-5 for citation-form signs, and 80.4% Top-1 and 93.0% Top-5 for signs pre-segmented from continuous signing, when using the combined isolated and pre-segmented sign datasets for training.

Performance enhancements are achieved through use of a bidirectional approach to harness the full temporal context of sign videos; and, for isolated sign clips, of both a Gating module, to filter out non-informative frames and an auxiliary multimodal branch for temporal action localization, to identify the start and end frames of signs. Temporal action localization is a critical step towards ASL recognition from fluent signing.

**7. Acknowledgments**

We are grateful to the many people who have helped with the collection, linguistic annotation, and sharing of the ASL data upon which we have relied for this research. In particular, we are indebted to the many ASL signers who have contributed to our database; to Gregory Dimitriadis at the Rutgers Laboratory for Computer Science Research, the principal developer of SignStream®, our software for linguistic annotation of video data (<https://www.bu.edu/asllrp/SignStream/3/>); to Matt Huenerfauth and his team for data collection at RIT; to DawnSignPress for sharing video data; to the many who have helped with linguistic annotations (especially Carey Ballard and Indya Oliver); and to Augustine Opoku, for development and maintenance of our Web-based database system for providing access to the linguistically annotated video data (<https://dai.cs.rutgers.edu/dai/s/dai>). This work was supported in part by NSF grants #2235405, #2212302, #2212301, and #2212303, but any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.



## 8. Bibliographical References

- Purva C Badhe and Vaishali Kulkarni. 2015. [Indian sign language translator using gesture recognition algorithm](#). In *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pages 195–200. IEEE.
- Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. 2021. [JOLO-GCN: Mining joint-centered light-weight information for skeleton-based action recognition](#). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2735–2744.
- Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. 2018. [Memory matching networks for one-shot image recognition](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4080–4088.
- João Carreira and Andrew Zisserman. 2017. [Quo vadis, action recognition? A new model and the Kinetics dataset](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.
- HM Cooper, Eng-Jon Ong, Nicolas Pugeault, and Richard Bowden. 2012. Sign language recognition using sub-units. *Journal of Machine Learning Research*, 13:2205–2231.
- Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitris Metaxas. 2022a. [Bidirectional skeleton-based isolated sign recognition using graph convolution networks and transfer learning](#). In *13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 7328–7338, Marseille, France. European Language Resources Association (ELRA).
- Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitris Metaxas. 2022b. [Isolated sign recognition using ASL datasets with consistent text-based gloss labeling and curriculum learning](#). In *Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives*, pages 13–20, Marseille, France. European Language Resources Association (ELRA).
- Nasser H Dardas and Nicolas D Georganas. 2011. [Real-time Hand Gesture Detection and Recognition using Bag-of-Features and Support Vector Machine Techniques](#). *IEEE Transactions on Instrumentation and Measurement*, 60(11):3592–3607.
- Cleison Correia de Amorim, David Macêdo, and Cleber Zanchettin. 2019. [Spatial-temporal graph convolutional networks for sign language recognition](#). In *International Conference on Artificial Neural Networks*, pages 646–657. Springer.
- Mark Dilsizian, Polina Yanovich, Shu Wang, Carol Neidle, and Dimitris Metaxas. 2014. [A new framework for sign language recognition based on 3D handshape identification and linguistic modeling](#). In *9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1924–1929, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. [RMPE: Regional Multi-person Pose Estimation](#). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. [Model-agnostic meta-learning for fast adaptation of deep networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1126–1135.
- Victor Garcia and Joan Bruna. 2017. [Few-shot learning with graph neural networks](#). *arXiv preprint arXiv:1711.04043*.
- Spyros Gidaris and Nikos Komodakis. 2018. [Dynamic few-shot visual learning without forgetting](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4375.
- Spyros Gidaris and Nikos Komodakis. 2019. [Generating classification weights with gnn denoising autoencoders for few-shot learning](#). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–30.
- Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard E Turner. 2018. [Meta-learning probabilistic inference for prediction](#). *arXiv preprint arXiv:1805.09921*.
- Kirsti Grobel and Marcell Assan. 1997. Isolated sign language recognition using Hidden Markov Models. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, volume 1, pages 162–167. IEEE.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *arXiv preprint arXiv:1503.02531*.

- Sepp Hochreiter and Jürgen Schmidhuber. 1997. [Long Short-Term Memory](#). *Neural Computation*, 9(8):1735–1780.
- Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. [3D Convolutional Neural Networks for Human Action Recognition](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.
- Songyao Jiang, Bin Sun, Lichen Wang, Yue Bai, Kunpeng Li, and Yun Fu. 2021. [Sign language recognition via skeleton-aware multi-model ensemble](#). *arXiv preprint arXiv:2110.06161*.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. [Imagenet classification with deep convolutional neural networks](#). In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data.
- Dongxu Li, Cristian Rodriguez Opazo, Xin Yu, and Hongdong Li. 2020. [Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison](#). In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1448–1458.
- Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019a. [Actional-structural graph convolutional networks for skeleton-based action recognition](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3595–3603.
- Wenbin Li, Jinglin Xu, Jing Huo, Lei Wang, Yang Gao, and Jiebo Luo. 2019b. [Distribution consistency based covariance metric networks for few-shot learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8642–8649.
- Frank R Lin, John K Niparko, and Luigi Ferrucci. 2011. [Hearing loss prevalence in the United States](#). *Archives of Internal Medicine*, 171(20):1851–1853.
- Hong Liu, Juanhui Tu, and Mengyuan Liu. 2017. [Two-stream 3D Convolutional Neural Network for skeleton-based action recognition](#). *arXiv preprint arXiv:1705.08106*.
- Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. 2018. [Learning to propagate labels: Transductive propagation network for few-shot learning](#). *arXiv preprint arXiv:1805.10002*.
- Dennis Looney and Natalia Lusin. 2021. [Enrollments in Languages other than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016 report](#). In *Modern Language Association*.
- Abbas Memiş and Songül Albayrak. 2013. [A Kinect Based Sign Language Recognition System using Spatio-temporal Features](#). In *Sixth International Conference on Machine Vision (ICMV 2013)*, volume 9067, pages 179–183. SPIE.
- Dimitris Metaxas, Mark Dilsizian, and Carol Neidle. 2018. [Linguistically-driven framework for computationally efficient and scalable sign recognition](#). In *11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1711–1718, Miyazaki, Japan. European Language Resources Association (ELRA).
- Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. 2017. [A simple neural attentive meta-learner](#). *arXiv preprint arXiv:1707.03141*.
- Ross E Mitchell, Travas A Young, Bellamie Bachelda, and Michael A Karchmer. 2006. How many people use ASL in the United States? why estimates need updating. *Sign Language Studies*, 6(3):306–335.
- Carol Neidle. 2023. [Challenges for Linguistically-driven Computer-based Sign Recognition from Continuous Signing for American Sign Language](#). *arXiv preprint arXiv:2311.00762*, pages 1–32.
- Carol Neidle and Carey Ballard. 2022. [Why alternative gloss labels will increase the value of the wlasl dataset](#). Report no. 21, American Sign Language Linguistic Research Project.
- Carol Neidle and Dimitris Metaxas. 2023a. ASLLRP Continuous Signing Corpora, version 1. <https://dai.cs.rutgers.edu/dai/s/signbank>. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.
- Carol Neidle and Dimitris Metaxas. 2023b. Boston University American Sign Language Lexicon Video Dataset (ASLLVD), version 7. <https://dai.cs.rutgers.edu/dai/s/signbank>. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.
- Carol Neidle and Dimitris Metaxas. 2023c. Dawn-SignPress (DSP) Collection, version 1. <https://dai.cs.rutgers.edu/dai/s/signbank>. American Sign Language Linguistic Research

- Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.
- Carol Neidle and Dimitris Metaxas. 2023d. Dawn-SignPress (DSP) Sentences Collection, version 2. <https://dai.cs.rutgers.edu/dai/s/signbank>. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.
- Carol Neidle and Dimitris Metaxas. 2023e. Rochester Institute of Technology (RIT) Collection, version 4. <https://dai.cs.rutgers.edu/dai/s/signbank>. American Sign Language Linguistic Research Project (ASLLRP) Sign Bank ©2022-2024, Boston and Rutgers Universities.
- Carol Neidle, Augustine Opoku, Carey M Ballard, Konstantinos M Dafnis, Evgenia Chroni, and Dimitris Metaxas. 2022a. Resources for computer-based sign recognition from video, and the criticality of consistency of gloss labeling across multiple large ASL video corpora. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 165–172, Marseille, France. European Language Resources Association (ELRA).
- Carol Neidle, Augustine Opoku, and Dimitris Metaxas. 2022b. ASL video corpora & Sign Bank: Resources available through the American Sign Language Linguistic Research Project (ASLLRP). *arXiv preprint arXiv:2201.07899*, pages 1–20.
- Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan Yuille. 2018. Few-shot image recognition by predicting parameters from activations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7229–7238.
- Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. ZS-SLR: Zero-Shot Sign Language Recognition from RGB-D Videos. *arXiv preprint arXiv:2108.10059*.
- Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International conference on learning representations*.
- Marcelo Sandoval-Castañeda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.
- Zed Sevcikova Sehyr, Naomi Caselli, Ariel M Cohen-Goldberg, and Karen Emmorey. 2021. The ASL-LEX 2.0 project: A database of lexical and phonological properties for 2,723 signs in American Sign Language. *Journal of Deaf Studies and Deaf Education*, 26(2):263–277.
- Bowen Shi, Diane Brentari, Greg Shakhnarovic, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. *arXiv preprint arXiv:2205.1287*.
- Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019a. Skeleton-based action recognition with directed graph neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7904–7913.
- Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019b. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVPR Conference on Computer Vision and Pattern Recognition*, pages 12026–12035.
- Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, 29:9532–9545.
- Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Tae Soo Kim and Austin Reiter. 2017. Interpretable 3D human action analysis with temporal convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 20–28.
- Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208.
- Sandrine Tornay, Oya Aran, and Mathew Magimai Doss. 2020. An HMM approach with inherent model selection for sign language and gesture recognition. In *12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 6049–6056, Marseille, France. European Language Resources Association (ELRA).

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Advances in Neural Information Processing Systems*, 30.
- Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. [Matching networks for one shot learning](#). 29.
- Ulrich von Agris, Moritz Knorr, and Karl-Friedrich Kraiss. 2008. [The significance of facial features for automatic sign language recognition](#). In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–6.
- Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. [Masked feature prediction for self-supervised visual pre-training](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658.
- Qinkun Xiao, Lu Li, and Yilin Zhu. [Skeleton-based few-shot sign language recognition](#). Available at *SSRN 4334054*.
- Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu. 2015. [Joint action recognition and pose estimation from video](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301.
- Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. [Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Quan Yang. 2010. [Chinese sign language recognition based on video sequence appearance modeling](#). In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1537–1542. IEEE.
- Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. [Temporal relational reasoning in videos](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818.





# Author Index

- Adam, Robert, 26  
Alba-Castro, José Luis, 386  
Andersen, Jari I., 290  
Asthana, Anushka, 102
- Balkstam, Eira, 254  
Battisti, Alessia, 1, 13  
Beautemps, Denis, 95  
Belleman, Robert G., 290  
Beskow, Jonas, 219  
Bigeard, Sam, 343  
Björkstrand, Thomas, 254, 343  
Bleicken, Julian, 184  
Bono, Mayumi, 26  
Borneman, Joshua, 213  
Börstell, Carl, 36, 46  
Böse, Oliver, 184  
Braffort, Annelies, 95, 204  
Bulla, Jan, 168
- Cabeza-Pereiro, Carmen, 386  
Challant, Camille, 354  
Chen, Yuxiao, 408  
Cho, Sukmin, 323  
Condé, Sther, 302  
Curiel, Arturo, 225
- Danet, Claire, 204  
De Meulder, Maartje, 54  
de Quadros, Ronice M., 282, 302  
de Vos, Connie, 168  
Desai, Aashaka, 54  
Docío-Fernández, Laura, 386
- Ebling, Sarah, 1, 13  
Efthimiou, Eleni, 244, 276, 343  
Esselink, Lyke D., 66
- Fabre, Diandra, 95  
Fernandes, Francisco, 282, 302  
Filhol, Michael, 77, 235, 354  
Fotinea, Stavroula-Evita, 244, 276, 343  
França, Diego, 282
- Gao, Yiran, 102
- Gavrilescu, Robert, 86  
Geißler, Thomas, 282  
Geraci, Carlo, 86  
Gibet, Sylvie, 315  
Gierman, Lisa, 178  
Göhring, Anne, 13  
Gökgöz, Kadir, 335  
Gouiffès, Michèle, 95, 204  
Goulas, Theodore, 276, 343  
Gurbuz, Sevgi, 213
- Halbout, Julie, 95  
Hall, Kathleen Currie, 102  
Han, Ligong, 395  
Hanke, Thomas, 184, 194, 276, 343  
Hansson, Patrick, 254  
Hara, Daisuke, 123  
Hobby, Grace, 102  
Hochgesang, Julie A., 54  
Holzknecht, Franz, 13  
Huerta-Enochian, Mathew John, 147  
Hwang, Eui Jun, 323
- Imashev, Alfarabi, 111  
Inoue, Jundai, 123  
Isard, Amy, 131, 184  
Islam, Shynggys, 111  
Israilov, Khassan, 111  
Itoyama, Katsutoshi, 370
- Kaneko, Hiroyuki, 262, 376  
Khan, Hafiz Muhammad Sarmad, 140  
Kim, Jung-Ho, 147  
Kimmelman, Vadim, 159, 168, 361  
Klezovich, Anna, 219  
Klomp, Ulrika, 178, 269  
Ko, Changyong, 147  
Ko, Seung Yong, 147  
Kocab, Annemarie, 54  
König, Lutz, 184  
Konrad, Reiner, 184  
Kopf, Maria, 276, 343  
Kuder, Anna, 343  
Kydyrbekova, Aigerim, 111

Langer, Gabriele, 194  
Lascar, Julie, 95, 204  
Le Naour, Thibaut, 315  
Lee, Huije, 323  
Loio, Milene Peixer, 282  
Lu, Alex X., 54

Makazhanov, Aibek, 111  
Malaia, Evie A., 213  
Malmberg, Fredrik, 219  
Manders, Pieter, 178  
Martínez-Guevara, Niels, 225  
Martinod, Emmanuella, 235  
McDonald, John, 244  
McLoughlin, Simon D., 140  
Mesch, Johanna, 86, 219, 254, 343  
Metaxas, Dimitris N., 395, 408  
Miwa, Makoto, 123  
Miyazaki, Taro, 262, 370, 376  
Mukushev, Medet, 111  
Müller, Anke, 194  
Murtagh, Irene, 140

Nakadai, Kazuhiro, 370  
Nauta, Ellen, 178  
Neidle, Carol, 395, 408

Okada, Tomohiro, 26  
Omardeen, Rehana, 276  
Oomen, Marloes, 66, 159, 178  
Otte, Felicitas, 194  
Otterspeer, Gomèr, 178, 269, 290  
Ouakrim, Yanis, 95

Park, Jong C., 323  
Pelupessy, Ray, 178  
Pérez-Pérez, Ania, 386  
Peters, Christian, 282  
Pfau, Roland, 159  
Picron, Frankie, 276  
Price, Ari, 168

Ranum, Oline A., 290  
Rathmann, Christian, 282, 302  
Reid, Maggie, 102  
Reverdy, Clément, 315  
Riemer Kankkonen, Nikolaus, 254  
Roelofsen, Floris, 66, 178, 269, 290  
Roh, Kyunggeun, 323  
Romanek, Peter Zalán, 302

Safar, Josefina, 168  
Şahin, Karahan, 335

Sandygulova, Anara, 111  
Sasaki, Yutaka, 123  
Schulder, Marc, 184, 343  
Sepke, Lea, 194  
Sharma, Paritosh, 354  
Sidler-Miserez, Sandra, 1  
Skobov, Victor, 26  
Stern, Galya, 178  
Susman, Margaux Marie Christelle, 361

Tan, Sihan, 262, 370  
Tismer, Christian, 276  
Tissi, Katja, 1  
Tkachman, Oksana, 102

Uchida, Tsubasa, 262, 376

Vacalopoulou, Anna, 343  
van den Bold, Emma, 13  
Van Landuyt, Davy, 276  
Vasilaki, Kyriaki, 343  
Vázquez-Enríquez, Manuel, 386  
Venter, Dalene, 178  
Vesik, Kaili, 102  
von Ascheberg, Thomas, 77

Wähl, Sabrina, 194  
Wójcicka, Joanna, 343  
Wolfe, Rosalee, 244, 276  
Wubbolts, Casper, 178

Xia, Zhaoyang, 395, 408

Yessenbayev, Zhandos, 111

Zhou, Yang, 395, 408