Person Identification from Pose Estimates in Sign Language

Alessia Battisti[*], **Emma van den Bold**[*], **Anne Göhring**[*],
Franz Holzknecht[†], **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.holzknecht@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[1] communities around the world. There are an estimated 5,500 native signers/early learners[2] 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 [3] (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.

[1] 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).

[2] 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.

[3] 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. Raters 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 ±2.17). Pose sequences were generated from the front view of the segments using MediaPipe Holistic (Grishchenko and Bazarevsky, 2020).[4] 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?

[4] 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[5].

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.[6]

[5] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

[6] 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.

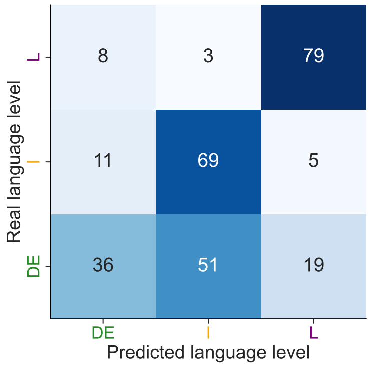


Figure 1: Confusion matrix for the language classifier predicting signers' language levels, evaluated using 10-fold cross-validation.

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

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	**23/60 (38.33%)**	**21/60 (35.0%)**	**9/40 (22.5%)**	**53/160 (33.12%)**
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%)
Total	53/310 (17.1%)	41/274 (14.96%)	23/274 (8.39%)	117/858 (13.64%)

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			**Signers I**			**Signers L**		
	1	2	**3**	4	**5**	6	**7**	**8**	**9**
Raters DE	7/20 (35.0%)	**9/20 (45.0%)**	7/20 (35.0%)	6/20 (30.0%)	11/20 (55.0%)	4/20 (20.0%)	**8/10 (80.0%)**	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%)	**14/19 (73.68%)**	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.

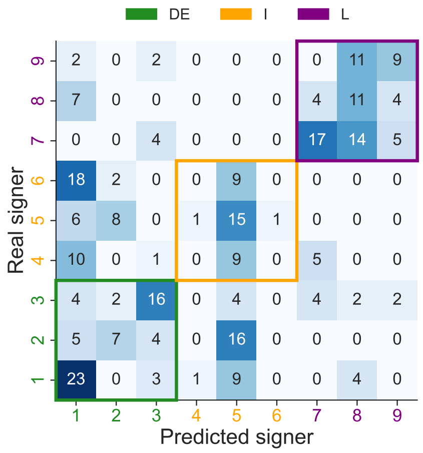


Figure 2: Confusion matrix for the signer classifier, evaluated using 10-fold cross-validation. The colored box indicates the language level of the signers.

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

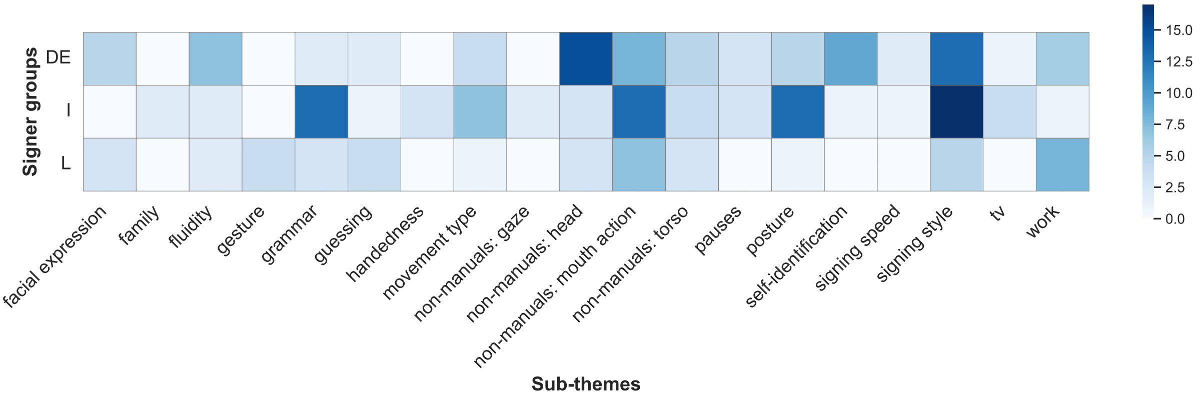


Figure 3: Matrix of the distribution of identifying factors across signer groups.

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 IICT (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 Braffort. 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: Realtime 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 Camgöz, 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. Sentence-bert: Sentence embeddings using siamese bert-networks.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Anonysign: Novel human appearance synthesis for sign language video anonymisation. 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 Schulder 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	mouth, gaze, eyebrows, head, torso
Signing	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 subthemes 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.

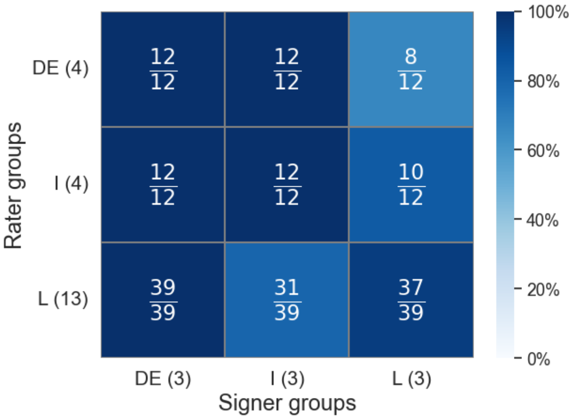


Figure C.2: Plot comparing the familiarity of signers across raters. Values in brackets indicate the number of persons in the group. Values within the cells denote the proportion of familiarity of familiarity between the raters and the signers, while the color gradient indicates the corresponding percentage.

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

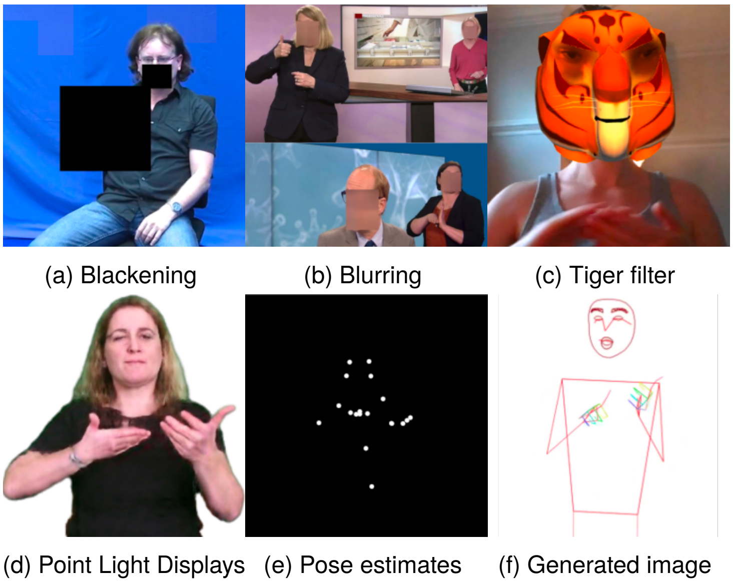
Table D.7: Average scores for the dummy multi-label signer classifier and multi-label signer classifier, evaluated with a 10-fold cross-validation.

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.



(a) Blackening (b) Blurring (c) Tiger filter

(d) Point Light Displays (e) Pose estimates (f) Generated image

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.

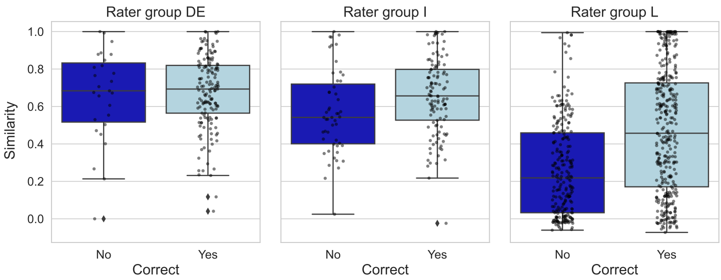


Figure E.3: Distribution of similarity scores for correct and incorrect identifications of the signers' language levels, across rater and signer groups.

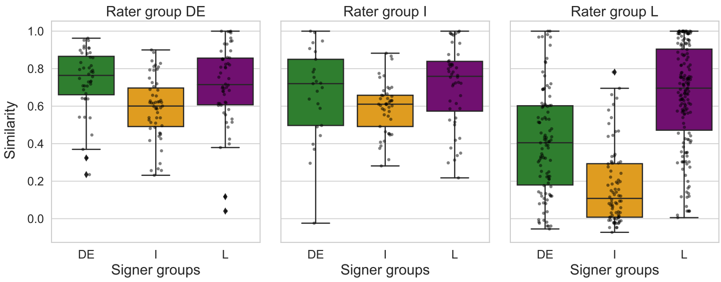


Figure E.4: Distribution of similarity scores for correctly identified language levels across signer groups. Each subplot corresponds to a different rater group and illustrates the distribution of similarity values (on the y-axis) obtained by rater groups in transcribing the content of the utterances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

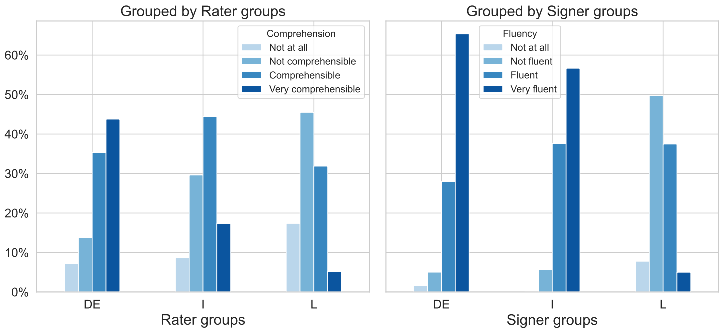


Figure E.5: Left: Bar plot showing the distribution of comprehension levels among rater groups. The y-axis represents percentages and the x-axis displays the four comprehension values across the rater groups. Right: Bar plot showing the distribution of fluency ratings among three signer groups. The y-axis 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*.

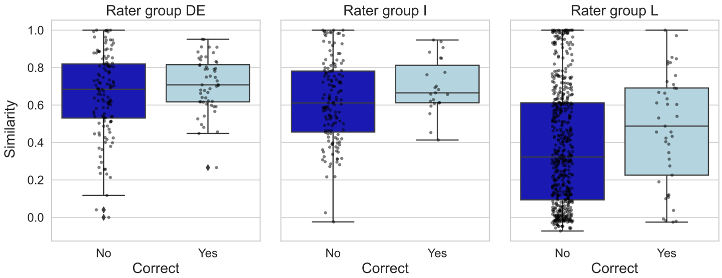


Figure F.6: Distribution of similarity scores for correct and incorrect signer identifications, across rater and signer groups.

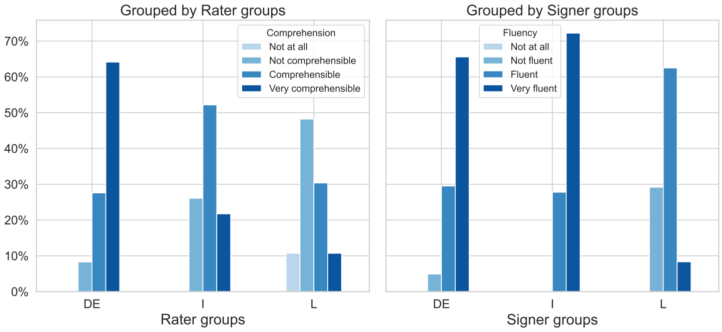


Figure F.7: Right: Bar plot showing the distribution of the comprehension ratings assigned by the raters to the stimuli whose signers were correctly identified. Left: Bar plot showing the distribution of fluency ratings among three signer groups.