

Workshop Proceedings

4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies

**Language Resources and Evaluation Conference (LREC)
Valletta, Malta, May 2010**

Workshop Organisers

Philippe Dreuw, RWTH, Aachen DE
Eleni Efthimiou, Institute for Language and Speech Processing, Athens GR
Thomas Hanke, Institute of German Sign Language, University of Hamburg, Hamburg DE
Trevor Johnston, Macquarie University, Sydney AU
Gregorio Martínez Ruiz, CRIC, Barcelona ES
Adam Schembri, Deafness Cognition and Language Research Centre, London GB

Programme Committee

Richard Bowden, University of Surrey, Guildford GB
Penny Boyes Braem, Center for Sign Language Research, Basel CH
Annelies Braffort, LIMSI/CNRS, Orsay FR
Onno Crasborn, Radboud University, Nijmegen NL
Patrice Dalle, IRIT, Toulouse FR
Evita Fotinea, Institute for Language and Speech Processing, Athens GR
John Glauert, University of East Anglia, Norwich GB
Jens Heßmann, University of Applied Sciences Magdeburg-Stendal, Magdeburg DE
Jette Kristoffersen, Professionshøjskolen UCC, Copenhagen DK
Lorraine Leeson, Trinity College, Dublin IE
Petros Maragos, National Technical University, Athens GR
Johanna Mesch, Stockholm University, Stockholm SE
Carol Neidle, Boston University, Boston US
Hermann Ney, RWTH, Aachen DE
Christian Rathmann, Institute of German Sign Language, Univ. of Hamburg, Hamburg DE
Antônio Carlos da Rocha Costa, Universidade Católica, Pelotas BR
Meike Vaupel, University of Applied Sciences Zwickau, Zwickau DE
Christian Vogler, Institute for Language and Speech Processing, Athens GR

Table of Contents

	Page
Editors' Preface	6
<i>Abdulaziz Almohimeed, Mike Wald, Robert Damper: An Arabic Sign Language Corpus for Instructional Language in School</i>	7
<i>Vassilis Athitsos, Carol Neidle, Stan Sclaroff, Joan Nash, Alexandra Stefan, Ashwin Thangali, Haijing Wang, Quan Yuan: Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms</i>	11
<i>Antonio Balvet: Issues Underlying a Common Sign Language Corpora Annotation Scheme</i>	15
<i>Nicola Bertoldi, Gabriele Tiotto, Paolo Prinetto, Elio Piccolo, Fabrizio Nunnari, Vincenzo Lombardo, Alessandro Mazzei, Rossana Damiano, Leonardo Lesmo, Andrea Del Principe: On the Creation and the Annotation of a Large-scale Italian-LIS Parallel Corpus</i>	19
<i>Roberto Borgotallo, Carmen Marino, Elio Piccolo, Paolo Prinetto, Gabriele Tiotto, Mauro Rossini: A Multilanguage Database for supporting Sign Language Translation and Synthesis</i>	23
<i>Helene Brashear, Zahoor Zafrulla, Thad Starner, Harley Hamilton, Peter Presti, Seungyon Lee: CopyCat: A Corpus for Verifying American Sign Language During Game Play by Deaf Children</i>	27
<i>Patrick Buehler, Mark Everingham, Andrew Zisserman: Exploiting Signed TV Broadcasts for Automatic Learning of British Sign Language</i>	33
<i>Pavel Campr, Marek Hruz, Jiří Langer, Jakub Kanis, Miloš Železný, Luděk Müller: Towards Czech On-line Sign Language Dictionary – Technological Overview and Data Collection</i>	41
<i>Anna Cavender, Neva Cherniavsky, Jaehong Chon, Richard Ladner, Eve Riskin, Rahul Vanam, Jacob Wobbrock: MobileASL: Overcoming the Technical Challenges of Mobile Video Conversation in Sign Language</i>	45
<i>Christophe Collet, Matilde Gonzalez, Fabien Milachon: Distributed System Architecture for Assisted Annotation of Sign Language Video Corpora</i>	49
<i>Genny Conte, Mirko Santoro, Carlo Geraci, Anna Cardinaletti: Why are you Raising your Eyebrows?</i>	53
<i>Helen Cooper, Richard Bowden: Sign Language Recognition using Linguistically Derived Sub-units.</i>	57
<i>Onno Crasborn, Han Sloetjes: Using ELAN for Annotating Sign Language Corpora in a Team Setting</i>	61
<i>Philippe Dreuw, Jens Forster, Yannick Gweth, Daniel Stein, Hermann Ney, Gregorio Martinez, Jaume Verges Llahi, Onno Crasborn, Ellen Ormel, Wei Du, Thomas Hoyoux, Justus Piater, Jose Miguel Moya Lazaro, Mark Wheatley: SignSpeak - Understanding, Recognition, and Translation of Sign Languages</i>	65
<i>Kyle Duarte, Sylvie Gibet: Corpus Design for Signing Avatars</i>	73
<i>Eleni Efthimiou, Stavroula-Evita Fotinea, Athanasia-Lida Dimou, Constandinos Kalimeris: Towards decoding Classifier function in GSL</i>	76
<i>Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, François Goudenove: DICTA-SIGN: Sign Language Recognition, Generation and Modelling with application in Deaf Communication</i>	80

<i>Ralph Elliott, Javier Bueno, Richard Kennaway, John Glauert: Towards the Integration of Synthetic SL Animation with Avatars into Corpus Annotation Tools</i>	84
<i>Michael Filhol, Maxime Delorme, Annelies Braffort: Combining constraint-based Models for Sign Language synthesis</i>	88
<i>Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, Hermann Ney: Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation</i>	92
<i>Carlo Geraci, Robert Bayley, Chiara Branchini, Anna Cardinaletti, Carlo Cecchetto, Caterina Donati, Serena Giudice, Emiliano Mereghetti, Fabio Poletti, Mirko Santoro, Sandro Zucchi: Building a Corpus for Italian Sign Language. Methodological Issues and some preliminary Results</i>	98
<i>Theodore Goulas, Stavroula-Evita Fotinea, Eleni Efthimiou, Michalis Pissaris: SiS-Builder: A Sign Synthesis Support Tool</i>	102
<i>Thomas Hanke, Lutz König, Sven Wagner, Silke Matthes: DGS Corpus & Dicta-Sign: The Hamburg Studio Setup</i>	106
<i>Thomas Hanke, Jakob Storz, Sven Wagner: iLex: Handling Multi-Camera Recordings</i>	110
<i>Julie A. Hochgesang, Pedro Pascual Villanueva, Gaurav Mathur, Diane Lillo-Martin: Building a Database while Considering Research Ethics in Sign Language Communities</i>	112
<i>Markus Hofmann, Kyle Goslin, Brian Nolan, Lorraine Leeson, Haaris Sheikh: Development of a Moodle VLE Plug-in to Support Simultaneous Visualisation of a Collection of Multi-Media Sign Language Objects</i>	116
<i>Matt Huenerfauth, Pengfei Lu: Eliciting Spatial Reference for a Motion-Capture Corpus of American Sign Language Discourse</i>	121
<i>Nedelina Ivanova: The Icelandic Sign Language Dictionary Project: Some Theoretical Issues</i>	125
<i>Tommi Jantunen: A comparison of two linguistic sign identification methods</i>	129
<i>Vince Jennings, Ralph Elliott, Richard Kennaway, John Glauert: Requirements for a Signing Avatar</i>	133
<i>Trevor Johnston: Adding Value to, and Extracting of Value from, a Signed Language Corpus through Secondary Processing: Implications for Annotation Schemas and Corpus Creation</i>	137
<i>Zdeněk Krňoul: New Features in Synthesis of Sign Language addressing Non-manual Component</i>	143
<i>Carlos R. Machado Oliveira: Adapting an Efficient Entry System for Sign Languages</i>	147
<i>Stefano Masneri, Oliver Schreer, Daniel Schneider, Sebastian Tschöpel, Rolf Bardeli, Stefan Bordag, Eric Auer, Han Sloetjes, Peter Wittenburg: Towards semi-automatic annotation of video and audio corpora</i>	150
<i>Guillem Massó, Toni Badia: Dealing with Sign Language Morphemes in Statistical Machine Translation</i>	154
<i>Silke Matthes, Thomas Hanke, Jakob Storz, Eleni Efthimiou, Nassia Dimiou, Panagiotis Kariotis, Annelies Braffort, Annick Choisier, Julia Pelhate, Eva Safar: Elicitation Tasks and Materials designed for Dicta-Sign's Multi-lingual Corpus</i>	158
<i>Nicholas Michael, Carol Neidle, Dimitris Metaxas: Computer-based Recognition of Facial Expressions in ASL: From Face Tracking to Linguistic Interpretation</i>	164

<i>Cedric Moreau, Bruno Mascret</i> : Data Organization in a Collaborative Sign Language Reference Tool	168
<i>Sara Morrissey, Harold Somers, Robert Smith, Shane Gilchrist, Sandipan Dandapat</i> : Building Sign Language Corpora for Use in Machine Translation	172
<i>Rie Nishio, Sung-Eun Hong, Susanne König, Reiner Konrad, Gabriele Langer, Thomas Hanke, Christian Rathmann</i> : Elicitation Methods in the DGS (German Sign Language) Corpus Project	178
<i>Ellen Ormel, Onno Crasborn, Els van der Kooij, Lianne van Dijken, Ellen Yassine Nauta, Jens Forster, Daniel Stein</i> : Glossing a Multi-purpose Sign Language Corpus	186
<i>Justus Piater, Thomas Hoyoux, Wei Du</i> : Video Analysis for Continuous Sign Language Recognition	192
<i>Vassilis Pitsikalis, Stavros Theodorakis, Petros Maragos</i> : Data-Driven Sub-Units and Modeling Structure of Multiple Cues for Continuous Sign Language Recognition	196
<i>Eva Safar, John Glauert</i> : Sign Language HPSG	204
<i>Rubén San-Segundo, Verónica López, Raquel Martín, David Sánchez, Adolfo García</i> : Language Resources for Spanish - Spanish Sign Language (LSE) Translation	208
<i>Adam Schembri, Onno Crasborn</i> : Issues in Creating Annotation Standards for Sign Language Description	212
<i>Jerry Schnepf, Rosalee Wolfe, John C. McDonald</i> : Synthetic Corpora: A Synergy of Linguistics and Computer Animation	217
<i>Marina Serrano, Jesús Gumiel, José M. Moya</i> : Automatic Sign Language Recognition and Translation: A social approach	221
<i>Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, Hermann Ney</i> : RWTH-Phoenix: Analysis of the German Sign Language Weather Forecast Corpus	225
<i>Saori Tanaka, Yosuke Matsusaka, Kaoru Nakazono</i> : Development of E-Learning Service of Computer Assisted Sign Language Learning: Online Version of CASLL	231
<i>Gudny Bjork Thorvaldsdottir</i> : You Get Out What You Put In: The Beginnings of Phonetic and Phonological Coding in the Signs of Ireland Digital Corpus	235
<i>Mara Vendrame, Gabriele Tiotto</i> : ATLAS Project: Forecast in Italian Sign Language and Annotation of Corpora	239
<i>Ulrich von Agris, Karl-Friedrich Kraiss</i> : SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition	243
<i>Alexander Voskresenskiy, Sergey Ilyin</i> : About Recognition of Sign Language Gestures	247
<i>Mark Wheatley, Annika Pabsch</i> : Sign Language in Europe	251
Author Index	255

Editors' Preface

This collection of papers stems from the Fourth Workshop on the Representation and Processing of Sign Languages, held in May 2010 as a satellite to the Language Resources and Evaluation Conference in Valletta, Malta.

While there has been occasional attention for sign languages at the main LREC conference, the main 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. For the second time, the workshop had sign language corpora as its main topic. With more time than in 2008, however, it was possible to include different views on corpora: The more linguistic aspects of collecting, annotating and analysing corpus data on the one hand, and sign language technologies, including vision and avatar technology, making use of sign language corpora or assisting the linguistic processing of corpora.

The papers at this workshop clearly identify the potentials of even closer cooperation between sign linguists and sign language engineers, and we think it is events like this that contribute a lot to a better understanding between researchers with completely different backgrounds.

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. We expect slides and posters to become available some time after the workshop at <http://www.sign-lang.uni-hamburg.de/lrec2010/programme.html>.

We would like to thank all members of the programme committee who helped us reviewing an unexpectedly high number of abstracts for the workshop within a very short timeframe!

Finally, we would like to point the reader to the proceedings of the previous workshops that form important resources in a growing field of research:

O. Streiter & C. Vettori (2004, Eds.) *From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication*.

[Proceedings of the Workshop on the Representation and Processing of Sign Languages. 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisbon.] Paris: ELRA.

C. Vettori (2006, Ed.) *Lexicographic Matters and Didactic Scenarios*. [Proceedings of the 2nd Workshop on the Representation and Processing of Sign Languages. 5th International Conference on Language Resources and Evaluation, LREC 2006, Genova.] Paris: ELRA.

O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitterlood (2008, Eds.) *Construction and Exploitation of Sign Language Corpora*. [Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation, LREC 2008, Marrakech.] Paris: ELRA.

(While the first two are available from ELRA, the third is available online at http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf.

For many presentations, slides or posters are also available at <http://www.sign-lang.uni-hamburg.de/lrec2008/programme.html>.)

The Editors

An Arabic Sign Language Corpus for Instructional Language in School

Abdulaziz Almohimeed, Mike Wald, Robert Damper

School of Electronics and Computer Science, University of Southampton
SO17 1BJ, UK

{aia07r|mw|rid}@ecs.soton.ac.uk

Abstract

An annotated sign language corpus is essential for machine translation projects. For this reason, many sign language corpora have been developed. Unfortunately, none of these is based on Arabic Sign Language (ArSL). In this paper, we present the ArSL corpus we created that is based on school-level language instruction.

1. Introduction

In recent years, many efficient machine translation approaches, both statistical and example-based, have been proposed. These are corpus-based approaches. The accuracy of translation is directly correlated to the size and coverage of the corpus. The corpus is a collection of translation examples constructed from existing documents, such as books and newspapers. A written system for sign language (SL) comparable to that used for natural language has not been developed. Hence, no SL documents exist, which complicates the procedure of constructing an SL corpus. In countries such as the UK, Ireland, and Germany, a number of corpora have already been developed and used for machine translation (MT). Unfortunately, there is no existing Arabic Sign Language (ArSL) corpus for MT. Therefore, a new ArSL corpus for language instruction was created.

2. Recent Work

The following is a survey of recent work that has informed our project.

The Centre for Deaf Studies in the School of Linguistics, Speech, and Communication Sciences, Trinity College Dublin built an Irish sign language corpus (Leeson et al., 2006). This corpus, which contains children's stories, took approximately three years to build. There were 40 signers involved. The participants' ages ranged from 18 to 65 years, and they came from different regions in Ireland. The recorded videos are about 20 hours long. The videos were annotated using the EUDICO Linguistic Annotator (ELAN)¹. The sign sentences were divided into different tiers that represent the Manual Features (MFs), referring to the hands, and Non-Manual Features (NMFs), referring to other parts of the body, such as the eyes, mouth, cheeks, etc., in gloss notation. In addition, an English translation was included for each sign sentence.

The European Cultural Heritage Online (ECHO) built a corpus of Swedish, British and Dutch SLs (Morrissey, 2008). It contains five children's stories signed in each SL. Approximately 500 signed sentences were collected in each language. ELAN was used to analyse the sentences.

Bungeroth et al. (2006) devised a German sign language corpus (DGS) for the weather report domain. They

constructed their corpus by extracting the German subtitle text and DGS translation from a German daily weather news television channel called Phoenix Broadcasts. The signs were collected by extracting the lower right corner of the broadcast frame that shows the DGS interpreter. They used ELAN to analyse the DGS sentences. They separated these sentences into the following five tiers: gloss notation of the sign sentences, word classes (such as verb, noun, adjective, adverb, etc.), DGS sentence boundaries, German sentence translation, and German sentence boundaries. There were 2,468 sentences collected. This corpus was mainly designed for statistical machine translation and sign recognition.

3. Gloss Notation

A gloss notation is a textual representation of sign language. It is beneficial to use this notation method because it allows for storing and processing the signs, and a sign avatar can represent and animate the signs by passing the details of MFs and NMFs. Arabic letters will be used for the ArSL corpus annotation. The reason for this is that none of the signers assisting with the corpus building has the ability to write the gloss in English. Therefore, a new specification for writing the gloss notation in Arabic has been created.

NMFs will now be used to describe the use of gloss notation. Each NMF is represented as follows:

```
(NMF Part) -- "Action" -- Action
                Description
```

Example: (Mouth) - "جاء" - شد

where "جاء" means the signer is pronouncing the word "جاء" (i.e., "jaa") and شد represents the signer stretching the lips. Table 1 summarises all of the gloss notations used for the ArSL corpus.

An example of this is the textual representation of the sign sentence of the Arabic sentence "السرقة حرام":

```
(Mouth) شد "جاء"
(Head)
(Eyes) اغلاق
(Nose)
```

The empty tiers mean no action exists. These annotation tiers can be combined as

```
(Mouth) شد "جاء" (Eyes) اغلاق
```

¹<http://www.lat-mpi.eu/tools/elan/>

Table 1: Summary of the Arabic gloss notation used in the ArSL corpus.

NMF	Action	Gloss
Eyes	Closing	“اغلاق”
	Opening	“فتح”
	Blinking	“ومض”
Nose	Wrinkling	“تجعد”
Mouth	Opening	“فتح”
	Closing	“اغلاق”
	Tongue out	“اخراج”
	Stre. lips	“شد”
	Sucking air	“شفط”
	Blowing air	“اخراج”
Shoulders	Forwards	“امام”
	Backwards	“خلف”
	Left	“يسار”
	Right	“يمين”
Cheeks	Puffing out	“مليء”
	Sucking in	“سحب”
Eyebrows	Raising	“اعلي”
	Lowering	“اسفل”

where the empty features will not be taken into account.

4. Corpus Setup

4.1. Domain

The translation system still needs a suitable dataset. By restricting the corpus domain, the input sentences can be covered by the matching corpus sentences, which will increase the accuracy of the translation results. In addition, since each word can have more than one meaning, depending on the context, a restricted domain will help reduce this ambiguity. The constructed corpus domain was restricted to the instructional language that is used in schools for deaf students. It can be described as a one-directional instruction that communicates sentences from teachers to students. For this purpose, a corpus team was established that included three native ArSL signers and one expert interpreter.

4.2. Video Recording

To be sure that the translated sign sentences are fluent, clear, complete, and fully independent from the original Arabic sentences, the recording steps in Figure 1 were followed.

In Figure 1, sign sentences were produced after the interpreter showed the signers the meaning of the sentences using ArSL, without having them read the Arabic sentence. The reason is that after reading the Arabic sentence, the signers signed all of the original Arabic sentence details, even if they were not required. The signers also followed the order of the Arabic sentence. Then they signed it. After each sentence was recorded, the video was checked by the native signers to be sure that it was correct and would be

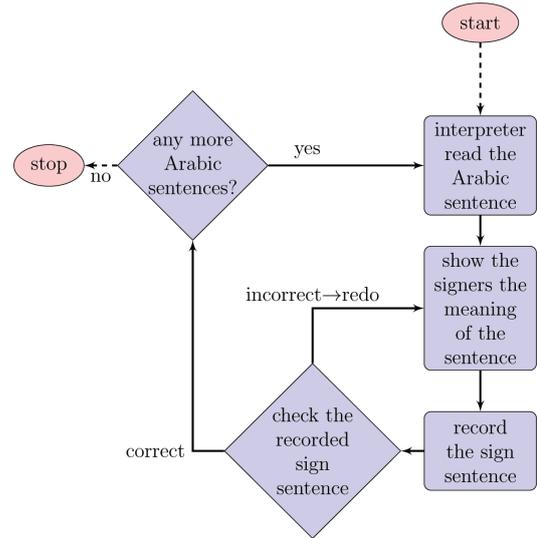


Figure 1: Steps in recording the signed sentences.

clear to deaf people in different age groups. If it was correct, they signed the next sentence; if not, the video was deleted and the sentence was recorded again. In the end, 213 ArSL sentences were recorded using a Sony DSC-W120 digital camera and were stored in MPEG format. The size of the recorded video frame is 640×480 pixels.

4.3. Video Corpus

After recording the sign sentences, the videos were annotated using the ELAN annotation tool. As shown in Figure 2, Arabic translation was added. Then, boundaries for each sign in the recorded video were clearly marked, and extra information was added. This information contained both MFs and NMFs. NMFs were described using the gloss notation discussed above. After isolating and adding the MFs and NMFs for all of the signs in the ArSL sentences, the annotated ArSL data were saved in EAF XML format.

4.4. Bilingual Corpus and ArSL Sign Dictionary

After the annotated ArSL data were saved in EAF XML format, the next phase was to build a bilingual corpus of ArSL and Arabic text delivered from the EAF and MPEG files. This procedure is essential for ArSL translation. The first step in constructing the bilingual corpus, is parsing the EAF XML files and extracting the MFs and NMFs for each sign, as shown in Figure 3.

Considering the information extracted for each feature (see bottom of Figure 3), the feature name field determines which part of the body is being used (this may be the right hand, left hand, mouth, etc.); the text shows the gloss notation for the particular body part. The EAF file name and Video fields identify the EAF and Video locations for each part. The start and finish time determines the exact location of the feature in the source video, which will be used later in constructing the signs-to-Arabic dictionary to extract the sign video clip from the source video. After the completion of this step, 1,897 features had been extracted. The next step in constructing the bilingual corpus is producing the sign dictionary using the extracted MFs

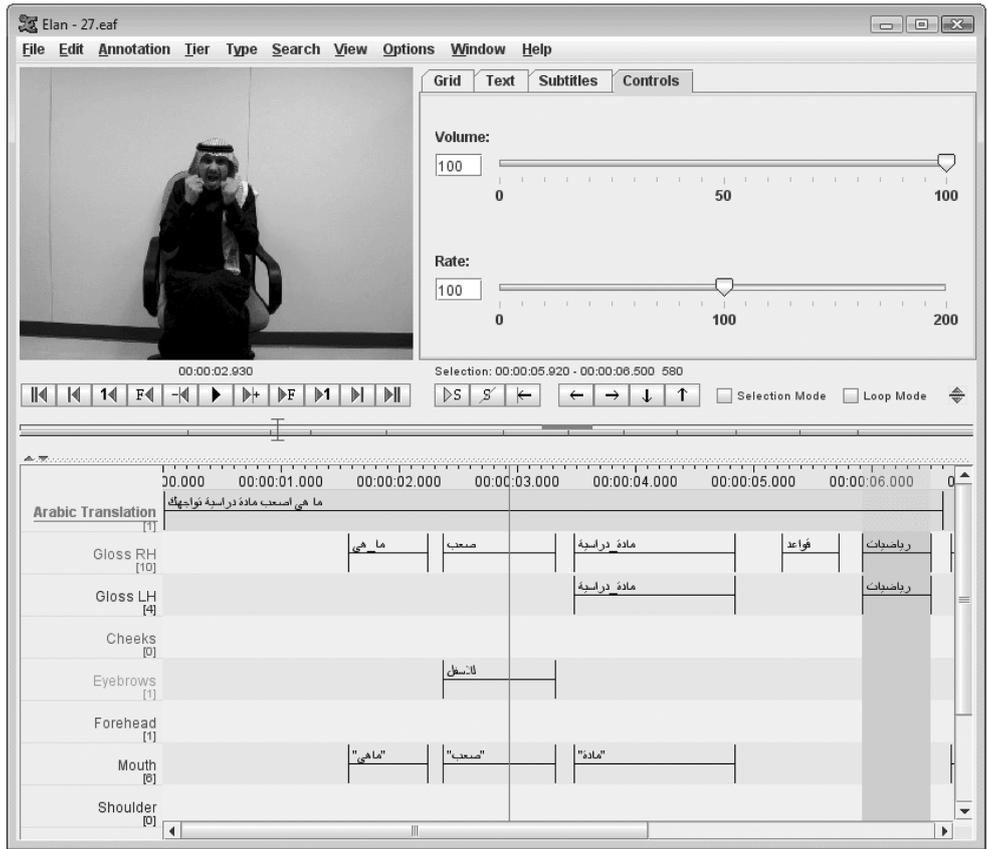


Figure 2: Using the ELAN Linguistic Annotator to annotate sign sentences.

```

<TIER TIER_ID="Arabic Translation" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a9" TIME_SLOT_REF1="ts1" TIME_SLOT_REF2="ts2"
    <ANNOTATION_VALUE>لا تضرب من هو اصغر منك</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
</TIER>
<TIER TIER_ID="Gloss RH" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a10" TIME_SLOT_REF1="ts3" TIME_SLOT_REF2="ts6">
    <ANNOTATION_VALUE>لا</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTATION_ID="a11" TIME_SLOT_REF1="ts9" TIME_SLOT_REF2="ts11">
  <ANNOTATION_VALUE>تضرب</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTATION_ID="a12" TIME_SLOT_REF1="ts13" TIME_SLOT_REF2="ts24">
  <ANNOTATION_VALUE>اصغر منك</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
</TIER>
<TIER TIER_ID="Gloss LH" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
<TIER TIER_ID="Cheeks" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
<TIER TIER_ID="Eyebrows" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a13" TIME_SLOT_REF1="ts4" TIME_SLOT_REF2="ts7">
    <ANNOTATION_VALUE>"تقوس للأعلى"</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
</TIER>
<TIER TIER_ID="Forehead" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
<TIER TIER_ID="Mouth" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a14" TIME_SLOT_REF1="ts5" TIME_SLOT_REF2="ts8">
    <ANNOTATION_VALUE>"لا"</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTATION_ID="a15" TIME_SLOT_REF1="ts5" TIME_SLOT_REF2="ts8">
  <ANNOTATION_VALUE>"تضرب"</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
  <ALIGNABLE_ANNOTATION ANNOTATION_ID="a16" TIME_SLOT_REF1="ts5" TIME_SLOT_REF2="ts8">
  <ANNOTATION_VALUE>"اصغر منك"</ANNOTATION_VALUE>
  </ALIGNABLE_ANNOTATION>
</TIER>
<TIER TIER_ID="Shoulder" LINGUISTIC_TYPE_REF="default-lt" DEFAULT_LOCALE="en">
  <ANNOTATION>
    <ALIGNABLE_ANNOTATION ANNOTATION_ID="a17" TIME_SLOT_REF1="ts5" TIME_SLOT_REF2="ts8">
    <ANNOTATION_VALUE>انحناء</ANNOTATION_VALUE>
    </ALIGNABLE_ANNOTATION>
  </ANNOTATION>
</TIER>

```

ID	eafFileName	VideoFileName	Part	aID	Stim	Ftim	txt
2175	D:\Corpus\EAF\9. MOV00011.MPG	Arabic Translation	a9	0	7524	لا تضرب من هو اصغر منك	
2176	D:\Corpus\EAF\9. MOV00011.MPG	Gloss RH	a10	1140	2320	لا	
2177	D:\Corpus\EAF\9. MOV00011.MPG	Gloss RH	a11	2490	3280	تضرب	
2178	D:\Corpus\EAF\9. MOV00011.MPG	Gloss RH	a12	4200	6100	اصغر منك	
2179	D:\Corpus\EAF\9. MOV00011.MPG	Eyebrows	a13	1140	2320	"تقوس للأعلى"	
2180	D:\Corpus\EAF\9. MOV00011.MPG	Mouth	a14	1140	2320	"لا"	
2181	D:\Corpus\EAF\9. MOV00011.MPG	Mouth	a15	2490	3280	"تضرب"	
2182	D:\Corpus\EAF\9. MOV00011.MPG	Mouth	a16	4200	6100	"اصغر منك"	
2183	D:\Corpus\EAF\9. MOV00011.MPG	Shoulder	a17	4390	5210	انحناء	

Figure 3: Parsing EAF XML.

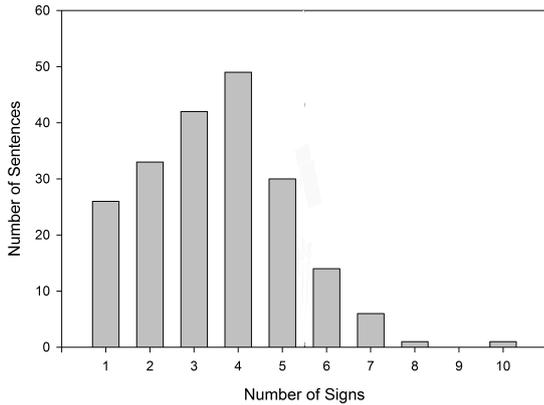


Figure 4: Distribution of collected sentences according to the number of signs that they contain.

and NMFs. All features that occur in the same period of time and have the same video source are considered to belong to same sign, and are collected together with that sign. Arabic sentence translation is then used to produce sign sentences in Arabic (bilingual corpus). The signs in this sentence will be linked to the corpus in the correct order using the video name.

The next step is extracting the video clips from the source video files using the start and finish times. After extracting each clip, the clip location will be appended to the sign table.

The last step is adding tags to represent the syntactic and morphological information for each sentence. The following is an example:

Arabic Sentence: لا تضرب من هو اصغر منك

After adding tags: <particle> لا

<verb present> تضرب

<preposition> من

<personal pronoun> هو

<adjective> اصغر

<preposition> من+

<personal pronoun> ك

In the end, there were 710 signs in the dictionary. There were 203 signed sentences in the bilingual corpus. The distribution of sentences according to the number of signs that they contain is shown in Figure 4.

5. Conclusion

We have presented an ArSL corpus for school-level language instruction. The corpus contains two main parts. The first part is the annotated video data that contains isolated signs with detailed information that includes MFs and NMFs. It also contains the Arabic translation script. The second part is the bilingual corpus that is delivered from the annotated video. A translation system can be used with a bilingual corpus. The ArSL corpus is now publicly available from www.ArSL.org and is suitable for ArSL recognition and translation systems. We are currently using the corpus to conduct translation experiments with Arabic text. We also plan to extend the number of examples to cover a larger domain.

Acknowledgments

This corpus could not have been constructed without the hard work of our expert signers: Ahmed Alzaharani, Kalwfah Alshehri, Abdulhadi Alharbi and Ali Alholafi.

6. References

- J. Bungeroth, D. Stein, P. Dreuw, M. Zahedi, and H. Ney. 2006. A german sign language corpus of the domain weather report. In *Fifth International Conference on Language Resources and Evaluation*, pages 2000–2003, Genoa, Italy.
- L. Leeson, J. Saeed, C. Leonard, and A. Macduff and D. Byrne-Dunne. 2006. Moving heads and moving hands: Developing a digital corpus of Irish sign language. ‘the signs of ireland’ corpus development project. In *Information Technology and Technology Conference 2006*, pages 33–43, Carlow, Ireland.
- S. Morrissey. 2008. *Data-Driven Machine Translation for Sign Languages*. Ph.D. thesis, School of Computer Science, Dublin City University, Ireland.

Large Lexicon Project: American Sign Language Video Corpus and Sign Language Indexing/Retrieval Algorithms

Vassilis Athitsos¹, Carol Neidle², Stan Sclaroff³, Joan Nash²,
Alexandra Stefan¹, Ashwin Thangali³, Haijing Wang¹, and Quan Yuan³

¹ Computer Science and Engineering Department, University of Texas at Arlington, Arlington, TX 76019, USA

² Linguistics Program, Boston University, Boston, MA 02215, USA

³ Computer Science Department, Boston University, Boston, MA 02215, USA

Abstract

Looking up the meaning of an unknown sign is not nearly so straightforward as looking up a word from a written language in a dictionary. This paper describes progress in an ongoing project to build a system that helps users look up the meaning of ASL signs. An important part of the project is building a video database with examples of a large number of signs. So far we have recorded video examples for almost all of the 3,000 signs contained in the Gallaudet dictionary (and some others not listed there). Locations of hands and the face have been manually annotated for a large number of videos. Using this data, we have built an application that lets the user submit a video of a sign as a query, and presents to the user the most similar signs from the system database. System performance has been evaluated in user-independent experiments with a system vocabulary of 921 signs. For 67% of the test signs, the correct sign is included in the 20 most similar signs retrieved by the system.

1. Introduction

Looking up the meaning of an unknown sign is not nearly so straightforward as looking up a word from a written language in a dictionary. This paper describes progress in an ongoing project to build a system that helps users look up the meaning of ASL signs. Our efforts in this project include construction of a large annotated video dataset, as well as system implementation.

Our dataset contains video examples for almost all of the 3,000 signs contained in the Gallaudet dictionary (and some others not listed there). Each video sequence is captured simultaneously from four different cameras, providing two frontal views, a side view, and a view zoomed in on the signer's face. Our video dataset is available on the Web.

In the current system, the user submits a video of the unknown sign to look up its meaning. The system evaluates the similarity between the query video and every sign video in the database, using the Dynamic Time Warping (DTW) distance. System performance has been evaluated in user-independent experiments with a system vocabulary of 921 signs. In our experiments we only use a single frontal view for both test and training examples. For 67% of the test signs, the correct sign is included in the 20 most similar signs retrieved by the system. More detailed results are presented in the experiments section.

Our approach is differentiated from prior approaches to sign language recognition by the fact that it is both vision-based and user-independent, while also employing a large vocabulary (921 signs). Many approaches are not vision-based, but instead use input from magnetic trackers and sensor gloves, e.g., (Gao et al., 2004; Vogler and Metaxas, 2003; Yao et al., 2006). Such methods have achieved good results on continuous Chinese Sign Language with vocabularies of about 5,000 signs (Gao et al., 2004; Yao et al., 2006).

On the other hand, computer vision-based methods typically have been evaluated on smaller vocabularies (20-250 signs) (Bauer and Kraiss, 2001; Deng and Tsui, 2002;

Dreuw and Ney, 2008; Fujimura and Liu, 2006; Kadir et al., 2004; Starner and Pentland, 1998; Zieren and Kraiss, 2005). While high recognition accuracy (85% to 99.3%) has been reported on vocabulary sizes of 164 signs (Kadir et al., 2004) and 232 signs (Zieren and Kraiss, 2005), those results are on user-dependent experiments, where the system is tested on users that have also provided the training data. In contrast, in our experiments the test signs are produced by users who do not appear in the training data, and the size of the vocabulary (921 distinct sign classes) is significantly larger than the vocabulary sizes that existing vision-based methods have been evaluated on.

2. Dataset: Videos and Annotations

In this section we describe the American Sign Language Lexicon Video Dataset (ASLLVD), which we have been building as part of this project. In particular, we update the information given in (Athitsos et al., 2008), to include the additional videos and annotations that we have added to this dataset in the last two years.

Our goal is to include video examples from a vocabulary that is similar in scale and scope to the set of lexical entries in existing ASL-to-English dictionaries, e.g., (Tennant and Brown, 1998; Valli, 2006). In the system vocabulary, we do not include name signs or fingerspelled signs, with the exception of some very commonly used ones (that are typically included in ASL dictionaries). We do not include classifier constructions, in which a classifier undergoes iconic movement, to illustrate the path or manner of motion, or the interaction of entities. The signs included in our dataset are restricted to the remaining (most prevalent) class of signs in ASL, which we refer to as "lexical signs."

At this point, we already have at least one video example per sign from a native signer, for almost all of the 3,000 signs contained in the Gallaudet dictionary (Valli, 2006). For a second signer we have collected 1630 signs, for a third signer we have collected 1490 signs, and for two additional signers we have collected about 400 signs. We would



Figure 1: One of the frontal views (left), the side view (middle), and the face view (right), for a frame of a video sequence in the ASL Lexicon Video Dataset. The frame is from a production of the sign “merry-go-round.”

eventually like to have at least three examples per sign for all signs in the system vocabulary.

2.1. Video Characteristics

The video sequences for this dataset are captured simultaneously from four different cameras, providing a side view, two frontal views, and a view zoomed in on the face of the signer. Figure 1 shows one of the frontal views, the side view, and the face view, for a frame of a video sequence in our dataset.

For the side view, the first frontal view, and the face view, video is captured at 60 frames per second, non-interlaced, at a resolution of 640x480 pixels per frame. For the second frontal view, video is captured at 30 frames per second, non-interlaced, at a resolution of 1600x1200 pixels per frame. All videos are available on the dataset websites, in formats employing both lossless compression (for higher video quality) and lossy compression (for faster downloading/browsing).

2.2. Annotations

The annotation for a video sequence contains, for each sign in that sequence, the start and end frames for that sign, a conventional English-based gloss of the sign, classification as one-handed or two-handed, and a signer ID. We also include manual annotations of the locations of the two hands and the face for a large number of signs. For hands, we mark at each frame the bounding box of the dominant hand, as well as the bounding box of the non-dominant hand for two-handed signs. For faces, we mark the bounding box of the face location at the first frame of each sign. Hand and face locations have been annotated for about 1500 sign examples from one signer, 1300 examples from a second signer, and 650 examples from a third signer.

The Gallaudet dictionary (Valli, 2006) includes a DVD containing a video example of every sign included in that dictionary. As those videos provide a valuable extra example per sign for almost all signs appearing in our dataset, we have annotated hand and face locations for about 1800 of the 3000 signs in that dictionary, and we intend to annotate the remaining signs in the next few months.

2.3. Availability

The ASLLVD dataset, including videos and annotations, is available for downloading on the project websites, located

at the following two URLs:

- http://csr.bu.edu/asl_lexicon
- http://vlml.uta.edu/~athitsos/asl_lexicon

In addition to the ASL Lexicon Video Dataset, a large quantity of ASL video and annotations that we have collected for previous projects is also available in various formats (on the Web from <http://www.bu.edu/asllrp/> and on CD-ROM; see also (Dreuw et al., 2008)). This video dataset includes 15 short narratives (2-6 minutes in length) plus hundreds of elicited sentences, for a total of about 2,000 utterances with over 1,700 distinct signs and a total of over 11,000 sign tokens altogether. These data have been annotated linguistically, using SignStream™ (Neidle, 2002; Neidle et al., 2001) (currently being reimplemented in Java with many new features). Annotations include information about the start and end point of each sign, part of speech, and linguistically significant facial expressions and head movements. The annotation conventions are documented (Neidle, 2002/2007) and the annotations are also available in XML format.

3. System Implementation

Signs are differentiated from one another by hand shape, orientation, location in the signing space relative to the body, and movement. In this paper we only use hand motion to discriminate between signs, leaving incorporation of hand appearance and body pose information as future work. Furthermore, we make the simplifying assumption that the system knows the location of the hands in all videos. The location of hands in all database sequences is manually annotated. Hand detection in the query sequence is performed in a semi-automatic way, where the system identifies hand locations using skin and motion information (Martin et al., 1998), and the user reviews and corrects the results.

Each sign video X is represented as a time series $(X_1, \dots, X_{|X|})$, where $|X|$ is the number of frames in the video. Each X_t , corresponding to frame t of the video, is a 2D vector storing the (x, y) position of the centroid of the dominant hand, for one-handed signs, or a 4D vector storing the centroids of both hands, for two-handed signs. For the purpose of measuring distance between the time-series representations of signs, we use the dynamic time warping (DTW) distance measure (Kruskal and Liberman, 1983).

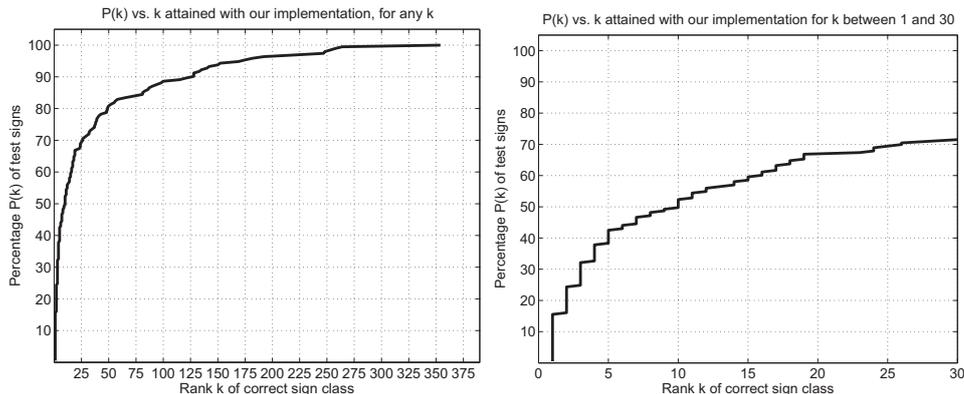


Figure 2: A plot of $P(k)$ vs. k illustrating the accuracy of our implementation. The x-axis corresponds to values of k . For each such value of k , we show the percentage of test signs $P(k)$ for which the correct sign class was ranked in the top k classes among all 921 classes. The plot on the right zooms in on a range of k from 1 to 30.

In particular, let Q be a test video and X be a training video. A warping path $W = ((w_{1,1}, w_{1,2}), \dots, (w_{|W|,1}, w_{|W|,2}))$ defines an alignment between Q and X . The i -th element of W is a pair $(w_{i,1}, w_{i,2})$ that specifies a correspondence between frame $Q_{w_{i,1}}$ of Q and frame $X_{w_{i,2}}$ of X . Warping path W must satisfy the following constraints:

- **Boundary conditions:** $w_{1,1} = w_{1,2} = 1, w_{|W|,1} = |Q|$ and $w_{|W|,2} = |X|$.
- **Monotonicity:** $w_{i+1,1} - w_{i,1} \geq 0, w_{i+1,2} - w_{i,2} \geq 0$.
- **Continuity:** $w_{i+1,1} - w_{i,1} \leq 1, w_{i+1,2} - w_{i,2} \leq 1$.

For one-handed signs, the cost $C(W)$ of the warping path W is the sum of the Euclidean distances between dominant hand centroids of corresponding frames $Q_{w_{i,1}}$ and $X_{w_{i,2}}$. For two-handed signs, we include in the cost $C(W)$ the sum of the Euclidean distances between non-dominant hands in corresponding frames. The DTW distance between Q and X is the cost of the lowest-cost warping path between Q and X , and is computed using dynamic programming (Kruskal and Liberman, 1983), with time complexity $O(|Q||X|)$.

To address differences in translation between sign examples, we normalize all hand centroid positions based on the location of the face. The face location in database videos is manually annotated, whereas for test videos we use the face detector developed by (Rowley et al., 1998). To address differences in scale, for each training example we generate 121 scaled copies. Each scaled copy is produced by choosing two scaling parameters S_x and S_y , that determine respectively how to scale along the x axis and the y axis. Each S_x and S_y can take 11 different values spaced uniformly between 0.9 and 1.1. We should note that each of these multiple copies is not a new sign video, but simply a new time series, and thus the storage space required for these multiple copies is not significant.

4. Experiments

The test set used in our experiments consists of 193 sign videos, with all signs performed by two native ASL sign-

ers. The training set contains 933 sign videos, corresponding to 921 unique sign classes, and performed by a native ASL signer different from the signers appearing in the test videos. When submitting a test sign, the user specifies whether that sign is one-handed or two-handed. The system uses that information to automatically eliminate from the results signs performed with a different number of hands (it should be noted, however, that, especially for certain signs, there can be some variability in the number of hands used). Although the ASLLVD dataset includes four camera views for each sign video, we only use the single 640x480 frontal view of each sign example in our experiments.

The results that we have obtained are shown in Figure 2. The measure of accuracy is a function $P(k)$ that measures the percentage of test signs for which the correct sign class was ranked in the top k out of the 921 classes. For example, in our results, $P(20) = 66.8\%$, meaning that for 66.8% of the 193 test signs, the correct sign class was ranked in the top 20 results retrieved by the system. In Figure 2, we include a plot focusing on a range of k from 1 to 30, as we believe few users would have the patience to browse through more than 30 results in order to find a video of the sign they are looking for. In Figure 3 we show an example of a query for which the correct match was ranked very low (rank 233), because of differences in the hand position between the query video and the matching database video.

On an Intel Xeon quad-core E5405 processor, running at 2.0GHz, and using only a single core, it takes on average 10 seconds to compute DTW distances and find the best matching results for a single test sign.

5. Discussion

In this paper we have provided an up-to-date description of the ASL Lexicon Video Dataset, a publicly available corpus that contains high-quality video sequences of thousands of distinct sign classes of American Sign Language, as well as manually annotated hand and face locations for a large number of those examples. We have also described an implementation of a system that allows users to look up the meaning of an ASL sign, with a simple method based on hand centroids and dynamic time warping.



Figure 3: Example of a query sign for which the correct class (“dog”) was ranked very low (rank 233). This sign exhibits small hand motion. A representative frame is shown for the query video (left) and for the correct database match (right). We note that the position of the hand is significantly different between the query and the database match.

Using our simple implementation, the correct class is ranked in the top 20 classes, out of 921 sign classes, for 67% of the test signs. This is an encouraging result, given that we are not yet using any information from handshape, hand orientation, or body pose. At the same time, our current implementation does not work very well for a significant fraction of test signs. For example, for 19% of the test signs the correct class is not included in the top 50. We hope that including additional information, from features related to hand and body pose, will lead to significantly better results, and that is a topic that we are currently investigating.

Acknowledgments

We gratefully acknowledge the following native signers, who have served as models for the project: Naomi Berlove, Elizabeth Cassidy, Lana Cook, Tyler Richard, and Dana Schlang. Annotations were carried out with the assistance of Jaimee DiMarco, Chrisann Papera, Jessica Scott, Jon Suen, and Iryna Zhuravlova. The research reported here has been partially funded by grants from the National Science Foundation: HCC-0705749, IIS-0812601, MRI-0923494.

6. References

- V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. 2008. The American Sign Language Lexicon Video Dataset. In *IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB)*.
- B. Bauer and K.F. Kraiss. 2001. Towards an automatic sign language recognition system using subunits. In *Gesture Workshop*, pages 64–75.
- J. Deng and H.-T. Tsui. 2002. A PCA/MDA scheme for hand posture recognition. In *Automatic Face and Gesture Recognition*, pages 294–299.
- P. Dreuw and H. Ney. 2008. Visual modeling and feature adaptation in sign language recognition. In *ITG Conference on Speech Communication*.
- P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *International Conference on Language Resources and Evaluation*.
- K. Fujimura and X. Liu. 2006. Sign recognition using depth image streams. In *Automatic Face and Gesture Recognition*, pages 381–386.
- W. Gao, G. Fang, D. Zhao, and Y. Chen. 2004. Transition movement models for large vocabulary continuous sign language recognition. In *Automatic Face and Gesture Recognition*, pages 553–558.
- T. Kadir, R. Bowden, E. Ong, and A. Zisserman. 2004. Minimal training, large lexicon, unconstrained sign language recognition. In *British Machine Vision Conference (BMVC)*, volume 2, pages 939–948.
- J. B. Kruskal and M. Liberman. 1983. The symmetric time warping algorithm: From continuous to discrete. In *Time Warps*. Addison-Wesley.
- J. Martin, V. Devin, and J.L. Crowley. 1998. Active hand tracking. In *Automatic Face and Gesture Recognition*, pages 573–578.
- C. Neidle, S. Sclaroff, and V. Athitsos. 2001. SignStreamTM: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320.
- C. Neidle. 2002. SignStreamTM: A database tool for research on visual-gestural language. *Journal of Sign Language and Linguistics*, 4(1/2):203–214.
- C. Neidle. 2002/2007. SignStream annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical report, American Sign Language Linguistic Research Project Nos. 11 and 13 (Addendum), Boston University. Also available at <http://www.bu.edu/asllrp/reports.html>.
- H.A. Rowley, S. Baluja, and T. Kanade. 1998. Rotation invariant neural network-based face detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 38–44.
- T. Starner and A. Pentland. 1998. Real-time American Sign Language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.
- R. A. Tennant and M. G. Brown. 1998. *The American Sign Language Handshape Dictionary*. Gallaudet U. Press, Washington, DC.
- C. Valli, editor. 2006. *The Gallaudet Dictionary of American Sign Language*. Gallaudet U. Press, Washington, DC.
- C. Vogler and D. N. Metaxas. 2003. Handshapes and movements: Multiple-channel American Sign Language recognition. In *Gesture Workshop*, pages 247–258.
- G. Yao, H. Yao, X. Liu, and F. Jiang. 2006. Real time large vocabulary continuous sign language recognition based on OP/Viterbi algorithm. In *International Conference on Pattern Recognition*, volume 3, pages 312–315.
- J. Zieren and K.-F. Kraiss. 2005. Robust person-independent visual sign language recognition. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, volume 1, pages 520–528.

Issues underlying a common Sign Language Corpora annotation scheme

Antonio Balvet

UMR 8163 STL (Université Lille 3 & CNRS),
 Université Lille Nord de France F-59653 Villeneuve d'Ascq,
 antonio.balvet@univ-lille3.fr

Abstract

Corpus-based Sign Language linguistics has emerged as a new linguistic domain, and as a consequence large-scale and controlled video data repositories are under construction for different Sign Languages. Nevertheless, as pointed by (Johnston, 2008) no unified annotation scheme is yet available, which compromises any chance of comparing or reusing corpora across research teams. Another related issue is the comparability of descriptions and formalizations between SL linguistics and mainstream linguistics. In this paper, we address the issue of the definition of a common annotation scheme for Sign Language corpora annotation, distribution, exchange and comparison. In section 2. we discuss the challenge of building inter-operable corpora for corpus-based linguistics. We also examine existing annotation schemes or strategies proposed for SL linguistics. In section 3. we propose a small set of annotation tiers, based on Frame-Semantics, as a common annotation scheme. We also propose to add text-level as well as utterance-level metadata to this common annotation scheme, in order to broaden the range of future uses of SL corpora.

1. Introduction

Mainstream corpus-based linguistics for oral and written languages is a flourishing research domain now that the capabilities of computers and linguistic software meet the demand of corpus-based and corpus-driven approaches both for linguistic research and applied domains of linguistics (second-language learning, lexicography, machine translation).

Sign Languages, on the other hand, are visuo-gestural and multi-segmental languages. Moreover, they have no stabilized written form, as of today, which hinders their computational processing. To make things even worse, Deafs over the world have generally been forbidden to use their natural language up until very recently¹, which has yielded great linguistic diversity. As a consequence, every aspect of their description, from the identification of basic units to the description of SL syntax or semantics, is a challenge to linguists, and even more so for computational or corpus linguists.

Sign Language linguistics can therefore be considered as a new and very challenging linguistic domain. Since most SL linguists are not native speakers of the language they are engaged in describing, at least some resort to actual language usage is necessary, even in the most formal approaches to SL linguistics. As a consequence, large-scale and controlled video data repositories are under construction for different Sign Languages: Auslan (Australian Sign Language), BSL² (British SL), DGS³ (German SL), LSF⁴ (French SL), and SSL⁵ (Swedish SL) to name but a few. The constitution of such controlled corpora is essential to the preservation and (formalized) description of Sign Languages in their diversity. Nevertheless, as pointed by (John-

ston, 2008) no unified annotation scheme is yet available, which compromises any chance of comparing or reusing corpora across research teams.

Another related issue is the comparability of descriptions and formalizations between SL linguistics and mainstream linguistics: given a set of SL corpora and their associated annotations, would a mainstream linguist be able to compare the syntax (or semantics, or any other traditional domain) of a given SL and the syntax of an oral language? Probably not, as most SL annotation schemes do not offer transcriptions (in their usual sense), and the glosses they provide are generally Sign-to-words intermediate associations rather than true morpheme-based interlinear glosses, as can be found in comparative linguistics and linguistic typology⁶.

In this paper, we address the issue of the definition of a common annotation scheme for Sign Language corpora annotation, distribution, exchange and comparison, focusing on some of the necessary features of such an annotation scheme, both from SL in general and from a computational (NLP or corpus-linguistics) perspective. In section 2. we discuss the challenge of building inter-operable corpora, for SL as well as mainstream corpus-based linguistics. We also examine an existing annotation scheme proposed for the Auslan project. In section 3. we propose a tentative common annotation schemes based on Frame-Semantics. We also propose to add text-level as well as utterance-level metadata to this common annotation scheme in order to broaden the range of future uses of SL corpora, with a computational perspective (corpus-linguistics and Natural Language Processing) in mind.

2. The challenge of corpus distribution, exchange and comparison

As stated above, SL linguistics has reached a crucial point: large-scale, controlled corpora are being devised all over

¹In the case of LSF, young Deafs were forbidden to sign during classes, up until 1991.

²<http://www.bslcorpusproject.org/>.

³<http://www.sign-lang.uni-hamburg.de/dgs-korpus/homee.html>.

⁴<http://www.creagest.cnrs.fr/>.

⁵<http://www.ling.su.se/pub/jsp/polopoly.jsp?d=12405&a=57659>

⁶See "The Leipzig Glossing Rules" for interlinear glosses examples: <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>.

the world, both for preservation and description objectives, mirroring a general trend in linguistics, which has caused large-scale, representative audio and text corpora to come to existence. Nevertheless, due to different practices, backgrounds (generative grammar vs. cognitive linguistics), funding opportunities, experimental setup (elicitation vs. free interaction, monologues versus dialogs), initial application (teaching SL vs. SL research), and also computer equipment and skills, no unified annotation scheme is yet available for all these projects, as pointed by (Johnston, 2008), which jeopardizes any chance of comparing or reusing corpora across research teams. This situation is not a privilege of SL research, though: it could be said that whenever two electronic corpora for any given (oral) language exist, they only seldom share the same tagsets, linguistic material, purposes, general methodology or even size. For example, if we consider two well-known English corpora such as the Penn Treebank (Marcus et al., 1993) and the Susanne corpus (Sampson, 1994), they vary wildly in coverage: over 1 million words for the Penn Treebank, versus around 130,000 for the Susanne corpus. They also vary wildly in their initial objectives: a large-scale “quasi-industrial” syntactically annotated corpus project for the Penn Treebank, versus a small-scale consistency-oriented project for the Susanne corpus. Of course, no common annotation scheme or even metadata exist for these corpora, which entails that each end-user should learn each corpus’s peculiarities. English is a well-established language, with a normalized written form and which has benefitted from a long grammatical history and an enduring research effort throughout the years. Therefore, corpus and computational linguists would be able to provide conversion tools whenever the need for inter-operability between the Penn Treebank and the Susanne corpus should arise. This is not the case for SLs in general: due to their multi-segmental and visuo-gestural modality, SLs have no normalized written form, which dramatically hinders their computational processing. At best, automatic recognition of only isolated parameters can be achieved, even with state-of-the art algorithms and pattern-recognition methods. Nevertheless, SL linguistics can benefit from the experience accumulated in mainstream corpus-linguistics. In our view, one way of guaranteeing SL corpora inter-operability are metadata.

2.1. Metadata: documenting and structuring corpora

Metadata can be considered as structured data on the data. In the framework of corpus linguistics, metadata generally serve two main purposes: The first one is the overall documentation of the source of the data, which generally entails identifying the speaker and his/her background (age, sex, education etc.), the interviewer or field linguist responsible for the data collection, the particular experimental setting used (types of cameras, exact reference, type of compression, type of recording medium: tapes, disks, flashdrives, use of lights, disposition of speakers, stimulus etc.), and other experimental variables. The second one is the structuring of each recorded corpus using *in situ* metadata so as to identify relevant discourse-level or utterance-level units (beginning and ending of a story, utterance or proposition boundaries, phonological/morphological/syntactic bound-

aries). For the purpose of corpus building, type 1 metadata are not necessarily included in the annotations associated with a given recording, while the latter generally are. Moreover, type 2 metadata are bordering on annotations, as the proper and consensual identification of many discourse or utterance-level units is a rather complex task. For example, even for written languages like English or French, the proper identification of such basic linguistic units as sentence boundaries or words is generally not an altogether easy task as inter- and even intraindividual variance are generally observed. In the domain of mainstream corpus-linguistics, the Text Encoding Initiative (TEI)⁷ offers guidelines and tools for the declaration of metadata (what to document) and the proper structuring of both overall metadata (type 1 above) and *in situ* metadata (type 2). In this framework, both discourse-level (text units) and utterance-level (sentences, words) units are identified, generally in order to support further annotations (e.g. lemmatization, part-of-speech tagging, syntactic parsing, semantic tagging). How are *in situ* metadata crucial for SL corpora? Because they provide the only proper (controlled) way, once a corpus is completely structured, to build sub-corpora out of the original corpus and the *in situ* metadata. In future uses of the SL corpora being devised to this date, we might want to consider cases where a researcher would need to study “the introduction of actants in stories told by left-handed Deaf children with a cochlear implant, from ages 5 to 7”. This would only be possible if such *in situ* metadata were included in the annotated files. To our knowledge, no SL annotation scheme allows for just such *in situ* labelling and subsequent potential selection of discourse as well as utterance-level units. Therefore, in our proposal for a common SL annotation scheme, we include metadata of the type discussed above: beginning and end of stories, utterances, propositions and possibly signed units.

2.2. A discussion of the Auslan annotation scheme/strategy

The Auslan project is a large corpus archive for Australian Sign Language: annotations are expected to take at least 10 years before they reach a stage compatible with extensive corpus-based research. To our knowledge, it is one of the only SL corpus annotation projects for which an annotation strategy has been explicitly devised and published, even though the same general approach can be found in other SL corpora projects, such as NGT. In the Auslan project, one of the solutions adopted by (Johnston, 2008) for consistent annotation relies on the concept of lemmatization, applied to Sign Language annotation: “the classification or identification of related forms under a single label or lemma (the equivalent of headwords or headsigns in a dictionary)”. Johnston describes the annotation protocol used for lexical signs in the framework of Auslan, where local interpretations of signs are normalized and constrained, in order to keep the set of lexical signs as small as possible: “[w]ithout lemmatization a collection of recordings [...] with various related annotation files [...] will not be able to be used as a true linguistic corpus as the counting, sorting, tagging.

⁷See TEI and TEI-Lite recommendations <http://www.tei-c.org/Guidelines/Customization/Lite/>

etc. of types and tokens is rendered virtually impossible.” This lemmatization process entails a high level of normalization and regularization, which in itself is not unusual in the course of corpus annotation. One of the key features of modern SL corpora, and more broadly of linguistic corpora in general, is their association with an annotation tool (Elan, Anvil, Transcriber, Praat, NiteXML...), which makes it possible to align annotations with the time indexes of the annotated media files (audio, video). Modern corpora are therefore associated with several time-aligned annotation layers, generally referred to as “tiers”. One of the most important feature of these annotation tiers is that they are not intended to preserve information (encode the original information in a different format), but rather to interpret and abstract over the original signal, in order to be integrated in a formalized description, and hopefully a model (a grammar) of the described language. Therefore, every time a linguistic corpus is built, annotation issues arise, requiring linguists to arrive at a compromise between faithfulness to the original data and consistency. As Johnston points out: “[w]ithout consistency (...) it will be impossible to use the corpus productively and much of the time spent on annotation will be effectively wasted because the corpus will cease to be, or never become, machine readable in any meaningful sense.”

3. Proposals for a common annotation scheme for lexical and non lexical signs

Lemmatization, or lexical sign normalization, appears as a necessary annotation strategy in the perspective of large and controlled SL corpora annotation. But, as (Johnston, 2008) points out: “[l]emmatization can only apply to lexical signs. However, many signed meaning units found in natural signed language texts are not lexical signs.” For Johnston “[lexical signs are] essentially, equivalent to the commonsense notion of *word*” whereas “the term *non-lexical sign* is reserved for a form that has little or no conventionalized or language-specific meaning value beyond that of its components in a given context.” Johnston proposes annotation conventions for such non lexical signs, of which the sub-category “depicting signs” seems to encompass what (Cuxac, 1996), and more specifically the Creagest team (Balvet et al., 2010), label **Highly Iconic Structures**. In the perspective of Cuxac’s semiological model of sign creation and development, these non lexical structures are a central linguistic device, both for natural human gestuality and Sign Languages. As Johnston’s citation above illustrates, this position is not shared by the vast majority of Sign Language linguists, who generally assume these structures to be peripheral at best, or even outside the range of language altogether (Garcia, 2010) and (Boutet et al., 2010).

Are lemmatas enough to ensure the linguistic exploitation and reusability of SL corpora among the SL linguistics community? Moreover, are lemmatas, in association with fine-grained postural and gestural descriptions, enough for ensuring comparability between SL and oral languages corpora? Could a mainstream linguist use SL annotations to compare structures among SLs and oral languages? Probably not, especially if one aims at describing not only lexical signs, but also Highly Iconic Structures which have been shown to represent over 40% of the semantic units in LSF

(and other LSs) stories and discourse⁸. Such structures are a major challenge for the formalized description of SLs: no oral language lemmatas are always available for each Transfer Structure, as they generally represent whole discourse units (propositions).

For all these reasons, we advocate in favor of Frame-Semantics primitives (Fillmore, 1977) and a Framenet-supported (Collin et al., 2008) annotation scheme for SL corpora. Frames are defined as “[having] many properties of stereotyped scenarios – situations in which speakers expect certain events to occur and states to obtain. In general, frames encode a certain amount of “real-world knowledge” in schematized form.” (Lowe et al., 1997). A typical example is the “commercial transaction” Frame, in which four Frame elements are generally required: two animated actants, an amount of money and an object. The result of the process associated with this Frame is the change of ownership of the object, in exchange for money. This stereotyped scenario can be associated with a relatively large set of lexical units in different languages (*buy, acheter, kaufen, comprar* etc.). Moreover, even though Frames are probably not universal concepts by essence, in our view they are likely to be learned and understood across different cultures and languages. And, as they represent basic stereotyped scenarios, they could be used to label complex Highly Iconic Structures, for which no direct mapping to a given oral language lemma can be found. Therefore, we feel that Frames are probably a useful tool for a common SL annotation scheme, not necessarily for glossing individual signed units, but at least as *in situ* metadata.

Therefore, we propose the following annotation tiers as a minimal common annotation scheme:

- text-level and utterance-level segments: START and END of stories, utterances, propositions;
- oral language glosses (e.g. English, French);
- Frame instance and core elements labels: Experiencer, Instrument, Goal, etc. based on the existing Framenet lexicon;
- lexical unit sets associated with Frame instances, as lemmatas for both lexical signs and Highly Iconic Structures.

To our knowledge, these annotations are not standard procedure in SL linguistics, except for glosses. Of course, they are not exclusive of finer-grained descriptions of phonological, morphological, syntactic or rhetorical constructs. But we believe this annotation strategy could overcome the limitations of resorting to lemmatas following Johnston’s annotation strategy. Moreover, including such Frame instances and core element labels could provide a common inter-operable indexing strategy, allowing researchers to extract comparable SL corpora segments based on their Frame instance labels, regardless of the particular sign languages or of the structures supporting the Frame instance (lexical sign, HIS).

⁸See (Sallandre, 2003) and (Cuxac and Sallandre, 2007).

In the figures below we give an example of the annotation of the sign GIVE⁹ and a Transfer Structure¹⁰ as instances of a GIVING Frame. In the LSF non lexical sign structure, signer Christelle signs “and then she gives her chicks a nice worm” using a complex Transfer Structure combining Situational Transfer (TREE) with Personal Transfer (signer = mother bird) and a clever adaptation of sign GIVE in order to resemble a beak configuration¹¹. This example is a clear instance of a whole proposition denoting a GIVING Frame, which cannot easily be mapped into lemma “GIVE”. It illustrates the necessity and usefulness of identifying such Frame instances, whether they are expressed with lexical signs or other more complex structures.



Figure 1: BSL Standard GIVE



Figure 2: LSF GIVE Transfer Structure

4. Conclusion and perspectives

In this paper, we have outlined a tentative common annotation strategy for SL corpora inspired by Frame-semantics, for the annotation of Frame instances, rather than just lemmatas, regardless of the particular SL or sign structure used.

⁹BSL, source: Spread The Sign web page, <http://www.spreadthesign.com>.

¹⁰LSF, source: LS-Colin corpus, see (Sallandre, 2003) for more details on the LS-Colin corpus.

¹¹See (Sallandre, 2003) for detailed transcriptions of HIS structures.

We believe this strategy could provide inter-operable SL corpora, which is crucial for their distribution, exchange and comparison. We include text-level and utterance-level metadata to our proposal, in order to broaden the future uses of the corpora being devised by allowing to derive narrower sub-corpora out of more generic ones.

5. 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. In *LREC (Language Resources and Evaluation Conference) 2010 Proceedings*, Malta.
- D. Boutet, M-A. Sallandre, and I. Fusellier-Souza. 2010. Gestualité humaine et langues de signes : entre continuum et variations. In B. Garcia and M. Derycke, editors, *Langage et Société*, number 131, pages 55–74. Maison des sciences de l’homme.
- C.F. Collin, C.J. Fillmore, and J.B. Lowe. 2008. The berkeley framenet project. In *Proceedings of the COLING-ACL*, pages 23–63.
- C. Cuxac and M-A. Sallandre. 2007. Iconicity and arbitrariness in French Sign Language: Highly Iconic Structures, degenerated iconicity and diagrammatic iconicity. In E. Pizzuto, P. Pietrandrea, and R. Simone, editors, *Verbal and Signed Languages - Comparing structures, constructs and methodologies*, pages 14–33. Mouton De Gruyter.
- C. Cuxac. 1996. *Fonctions et structures de l’iconicité. Analyse descriptive d’un idioloecte parisien de la Langue des Signes Française*. Ph.D. thesis, Université Paris 5.
- C.J. Fillmore. 1977. The need for a frame semantics in linguistics. *Statistical Methods in Linguistics*, (12):5–29.
- B. Garcia. 2010. *Sourds, surdit , langue(s) des signes et  pist mologie des sciences du langage. Probl matiques de la scripturisation et mod lisation des bas niveaux en Langue des Signes Fran aise (LSF)*. Habilitation thesis, Universit  Paris 8–Saint-Denis.
- T. Johnston. 2008. Corpus linguistics and signed languages: no lemmata, no corpus. In *LREC (Language Resources and Evaluation Conference) 2008, Proceedings of the workshop on the representation and processing of Sign Languages*, pages 82–87.
- J.B. Lowe, C.F. Baker, and C.J. Fillmore. 1997. A frame-semantic approach to semantic annotation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, D.C. SIGLEX.
- M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.
- M-A. Sallandre. 2003. *Les unit s du discours en Langue des Signes Fran aise (LSF). Tentative de cat gorisation dans le cadre d’une grammaire de l’iconicit *. Ph.D. thesis, Universit  Paris 8–Saint-Denis.
- G. Sampson. 1994. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Clarendon Press.

On the creation and the annotation of a large-scale Italian-LIS parallel corpus

**Nicola Bertoldi¹, Gabriele Tiotto², Paolo Prinetto², Elio Piccolo²,
Fabrizio Nunnari³, Vincenzo Lombardo³, Alessandro Mazzei⁴,
Rossana Damiano⁴, Leonardo Lesmo⁴, Andrea Del Principe⁵**

(1) Fondazione Bruno Kessler – Trento, Italy – bertoldi@fbk.eu

(2) Dip. di Automatica e Informatica, Politecnico di Torino – Turin, Italy – {gabriele.tiotto,paolo.prinetto,elio.piccolo}@polito.it

(3) Virtual Reality and Multimedia Park – Turin, Italy – {fabrizio.nunnari,vincenzo.lombardo}@vrmmp.it

(4) Dip. di Informatica, Università di Torino – Turin, Italy – {mazzei,damiano,lesmo}@di.unito.it

(5) Centro Ricerche ed Innovazione Tecnologica, RAI – Turin, Italy – andrea.delprincipe@rai.it

Abstract

This paper presents the current development of the first large parallel corpus between Italian and Italian Sign Language (Lingua Italiana dei Segni, LIS). This initiative has been taken within the ATLAS project (Automatic Translation into Sign Languages), that aims at realizing a virtual interpreter, which automatically translates an Italian text into LIS. The Italian-LIS virtual interpreter is implemented by means of two modules interfaced by the ATLAS Extended Written LIS (AEWLIS), which is a translation-oriented representation of LIS: the first module translates the source Italian text into AEWLIS; the second module transforms the AEWLIS content into a coherent LIS sequence, smoothly animated by a virtual character. As no significant amount of electronic data are available for Italian and LIS, we have started building a parallel corpus from scratch in order to train and tune the Italian-AEWLIS translation system, and to compare the resulting virtual animations with human-performed LIS interpretations. The corpus, which will be freely available, actually presents a tri-lingual structure, with the Italian text, the AEWLIS sequence, and the signed LIS video.

1. Introduction

People who were born deaf or acquired deafness in the first years of life -approximately 70,000 in Italy- experience big obstacles to integrate into the society, because they could not properly acquire knowledge of the spoken language, and consequently of the written language, and vice versa hearing people very rarely practice Sign Languages (SLs). The care of hearing-impaired people progressively grows; the increasing request for SL interpretation in educational, legal, and health contexts is foreseen and soon expected to be extended to culture and entertainments. The depicted scenario makes clear the relevance of the availability of a low cost technology to support the SL interpretation.

ATLAS (Automatic Translation into sign LAnguageS) is a three-year project, funded by the local government of Piedmont, Italy, aiming at providing Italian deaf people with facilities to access broadcast communications, and in particular to follow TV programmes. More specifically, ATLAS aims at developing a virtual interpreter, which automatically translates Italian into LIS.

The virtual interpreter has a modular structure and relies on a translation-oriented symbolic representation of the LIS, called ATLAS Extended Written LIS (AEWLIS). Training and tuning of most components of the virtual interpreter requires a parallel corpus, composed of a large set of Italian sentences, their human-performed LIS interpretations and their corresponding AEWLIS. Furthermore, an excerpt of this parallel corpus is exploited for the component-wise and end-to-end evaluation in terms of both automatic and subjective criteria.

As a significant amount of parallel data is not available yet, we have started building a new corpus from scratch. The corpus actually presents a tri-lingual structure, with the Italian text, the AEWLIS sequence, the signed LIS videos. The

first release of the corpus will contain weather forecast bulletins for a total of about 15K Italian running words and about 1.5 hours of LIS videos.

Next Section reports on scientific projects about the automatic translation of Sign Languages around the world. Section 3. briefly overviews the full-fledged virtual interpreter developed within the ATLAS project. Section 4. describes the corpus which the ATLAS partners are building and discusses issues arised during its creation. Section 5. presents AEWLIS, the intermediate artificial language chosen for representing the LIS in a written form. Some conclusions are finally drawn in Section 6..

2. State-of-the-art of the research on SLs

Since early 90's the scientific research on SLs has been constantly growing because of the increasing care of deaf people, their augmenting willingness of integration into the society, and the availability of more and more powerful computing facilities and software which make possible the automatic dealing with SL.

Most research projects on SLs approach American, British, Dutch, and German SLs: (SignWriting, 2009; DictaSign, 2010; eSign, 2009; SignSpeak, 2010; U.DePaul, 2008; U.Boston, 2002; Echo, 2010; NGTCorpus, 2009). They address some (or all) of the following highly-interconnected tasks: SL recognition, 3D character animation, machine translation, production of SL dictionaries and corpora, development of toolkits for SL annotation and transcription like (ELAN, 2010), and integration with existing communication devices, like mobile phones (mobileASL, 2010).

Differently from most oral languages, SLs do not have a natural corresponding written expression. Hence, researcher have proposed many artificial languages to represent them into a written form: gloss-based, (Stokoe et al.,

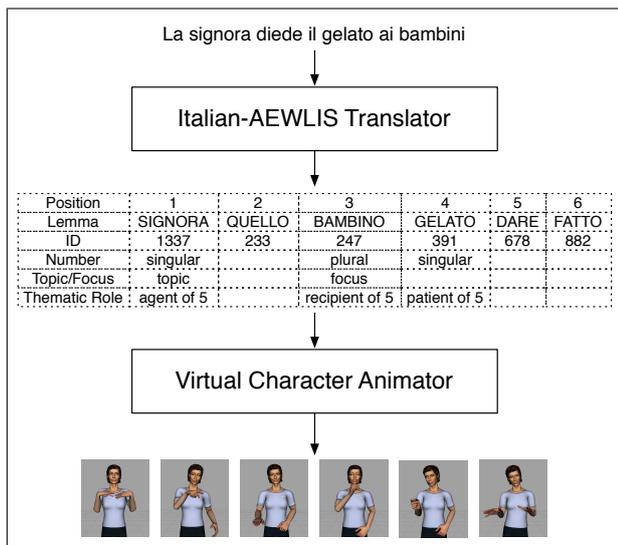


Figure 1: The interpretation process is performed by Donna, the ATLAS virtual interpreter, in two steps. First the Italian text is transformed/translated into AEWLIS, consisting of a sequence of signs and syntactic/semantic relationships among them. Then, a virtual character animates the AEWLIS content into a coherent LIS sequence.

1965), (HamNoSys, 2004), (SignWriting, 2009), (Elliott et al., 2004; Filhol and Braffort, 2006). Each of these transcription methods is tuned to answer the specific need they were developed for: understanding and translating sentence meaning, producing and recognizing isolated sign, producing digital animation, etc.

SLs present many phenomena requiring the adaptation of a sign to the context of the signed sentence: relocation in the signing space, increase or reduction of the “size”, repetition (e.g. for plural), movement of hands through the space from and to context-dependent positions (e.g. for verbs like “to go” or “to give”), the use of hand-shapes as classifiers (Huenerfauth, 2006). Recent research on 3D animation deals with these context-dependent phenomena; through synthetic approach the character is animated with an animation language which is interpreted by a real-time player (Veale et al., 1998; Marshall and Sáfár, 2003).

Both rule-, example-, and, more recently, statistical-based approaches have been adopted for the text-to-SL machine translation. Most MT systems exploit a gloss-based notation of SL. A detailed overview of SL MT can be found in (Morrissey, 2008).

Finally, several research projects focus on collecting dictionaries and corpora related to SLs: (Echo, 2010; eLIS, 2006; SignWriting, 2009; Bungeroth et al., 2006; Bungeroth et al., 2008; BSLCorpus, 2010; NGTCorpus, 2009).

As concerns the LIS, very few and small-size projects have been funded (eLIS, 2006; DizLis, 2010; BlueSign, 2010). The ATLAS project aims at covering this gap, fostering the interest of the research community towards LIS and LIS-related tools. It is worth remarking that all software and linguistic resources developed by ATLAS partners will be made freely available.

3. The ATLAS virtual interpreter

Figure 1 provides a high-level representation of the full-fledged Italian-LIS virtual interpreter. It is actually implemented by means of two modules interfaced by AEWLIS: the **Italian-AEWLIS Translator** and the **Virtual Character Animator**.

The **Translator** transfers the meaning of an Italian text into AEWLIS by means of both statistical and rule-based techniques. The phrase-based statistical Machine Translation (MT) system relies on the high-performing state-of-the-art toolkit Moses (Koehn et al., 2007). The rule-based MT system exploits linguistic rules to connect the morphology, syntax and semantics of the source language to those of the target language. The linguistic knowledge about Italian used by the translator has already been exploited (Lesmo et al., 2009) During the project, the integration of the two systems will be investigated.

The **Animator** relies on a *signary*, a repository in which each sign is described by an animation language in terms of motion data (motion-captured or hand-made), procedural animations and applicable parameters for size, relocation, repetition, hand-shapes, etc. First, a motion planner transforms the information present in the AEWLIS sequence into a sequence of signs, taken from the signary, whose parametric values are determined according to the actual context. Then, a blending system, a technique widely used in videogame architectures, creates the LIS by smoothly joining in real-time the existing animation clips through interpolation functions.

Animations will be displayed, for both broadcast and on-demand delivering, on a variety of user terminals (including DVB, Web, Mobile Phones). The heavy computational effort (translator and motion planner) will be carried out on a centralized server, and the visual rendering will be performed on the device. The physical appearance of the virtual character responds to criteria that enhance the perception of hand motion and facial expressions, that are fundamental in understanding signs. We have designed two signing characters, Donna and Manuel.

4. Description of the corpus

The module **Italian-AEWLIS Translator** introduced in Section 3. requires the availability of a parallel corpus for its training. As this kind of electronic data are not available for Italian yet, we have started building a parallel corpus from scratch.

The first application domain of the ATLAS project is the automatic interpretation of weather forecast bulletins daily broadcasted by RAI (Radio Televisione Italiana). Hence, we collected a set of 55 bulletins of 2008, containing about 15K Italian words corresponding to about 1.5 hours of Italian audio/video.

According to LIS experts and interpreters, we defined the following procedure to build the parallel corpus. First, the audio of the TV bulletin is automatically transcribed by a speech recognition system (Brugnara et al., 2000) and manually checked to correct transcription errors. Portions of the bulletin which is not strictly related to the weather domain are eventually removed. Then, a LIS expert interprets the content of the cleaned text and a movie of his/her signing

	Italian	AEWLIS
Number of bulletins		55
Number of sentences		585
Running terms	15,012	6,000*
Average terms per sentence	25.7	10*
Dictionary Size	1,442	300*
Singletons	614	-

Table 1: Statistics of the first release of the Italian-AEWLIS parallel corpus; asterisks mark estimated statistics.

is recorded using a standard framing. In order to avoid an unnecessary variability of the LIS, the expert is committed to sign a genuine but plain LIS. Finally, the same expert annotates his/her LIS movie according to the AEWLIS annotation guidelines described in Section 5..

As AEWLIS has several independent annotation levels (see Section 5.), they can be marked in successive steps from the least to the most specific. An editor has been developed by ATLAS partners to support the expert in the annotation process, which will become a Computer Assisted Translation tool after the integration with the **Translator**.

Thus, the corpus actually results in a tri-lingual structure, with the Italian text, the AEWLIS sequence, the signed LIS video. Furthermore, we would have also audio/video of the original TV bulletins and the automatic transcription, but at present these data are not exploited in the project.

The corpus is currently under development. we have completed about a third of the expected final size. The creation of the corpus will be presumably finalized by the end of June 2010, and made publicly available to the community. Portions of the corpus will be extracted to create development and test sets for tuning and evaluation purposes. Some statistics of the parallel corpus are reported in Table 1; asterisks mark estimated statistics based on the actual partial corpus. An example of an AEWLIS-annotated sentence is shown and commented in Figure 2. Of course, Italian audio/video and LIS movie are not reported. The AEWLIS annotation is reported here in a simplified human-readable format, and only relevant information are reported.

The generation of the parallel corpus is time-consuming and expensive, because an intensive effort of skilled human is required both for signing and annotating. In order to get the best trade-off between the size of the training corpus and the cost for collecting it, smart solutions have been adopted: split sentences into small segments conveying (self-)consistent content and syntax; avoid duplicate or highly overlapping segments; incrementally collect segments which are more distant (for instance, with respect to Levenshtein distance) from those already gathered.

The corpus domain has caused the creation of new, intuitively iconic, signs for the human interpreters, for a number of concepts that would have caused long boring paraphrases. This is common practice among LIS speakers; in particular, LIS interpreters need to agree on a limited number of novelties in order to keep the variation among interpreters at a minimum. The LIS signed and annotated in the corpus is not “spontaneous” like in a conversation, but “genuine” like that produced by professionals interpreting, for example, TV programmes.

5. ATLAS Extended Written LIS

AEWLIS is a formal language defined within the ATLAS project and plays a two-fold role: it provides a symbolic representation for the annotation of the LIS corpus, and it is the interchange language between the **Translator** and the **Animator**. AEWLIS format encompasses both functions, but different and possibly overlapping subsets are employed for the two tasks. AEWLIS is translation-oriented in the sense that it contains all information required to (i) convey the meaning of the original Italian sentence and (ii) “instruct” or “pilot” the virtual character to fluently sign it. We know that there is no consensus about the possibility to encode a signed language in a written form. Indeed, we do not assert that AEWLIS is a “linguistic” written form of LIS: AEWLIS contains the necessary (phonologic, syntactic, semantic) information that the virtual character needs to properly realize the LIS sequence.

The annotation is performed at a sentence level, so links to other elements of other sentences of the same (or different) bulletin(s) are not allowed. Each sentence is split into a sequence of Time Slices (TSs), each defined as the time interval needed to perform a sign. A TS is considered atomic in the sign sequence. It is worth noticing that in the annotation phase we do not actually perform a time segmentation of the LIS sequence, but we simply associate a TS with each single sign. Indeed, the goal of the project is not the development of a sign recognizer, which would probably rely on such information.

AEWLIS includes three main kinds of annotation. The first level describes the meaning conveyed by the actual sign assigning a Lemma (or gloss) to each TS. The syntactic number (singular/plural) is possibly reported. A Sign-ID identifies the sign in the signary, if any¹.

The second level independently describes all Communication Channels relevant in LIS: Left and Right Hands, Direction, Body, Shoulder, Head, Facial, Labial, Gaze. Practically, only the modifications with respect to the neutral default for the corresponding sign are annotated. Specific annotation for the Left and Right Hands is given if they realize distinct signs contemporarily².

The third level provides a shallow syntactic/semantic structure of the sentence if available, by reporting for each TS its parent and its role. The main thematic roles proposed in (Petukhova and Bunt, 2008) are reported which can have a strong impact on the animation: agent/patient, initial/final location, etc. Topic and Focus (Hajičová et al., 1998) (Lillo-Martin and de Quadros, 2004) of the sentence are possibly annotated; the speech act specifies whether the sentence is declarative, imperative or interrogative.

Furthermore, AEWLIS has been defined as a set of independent annotation levels (*Tags*), which can be filled at different moments, or even left empty. The only mandatory tag is the Lemma. All the annotation Tags are associated to a single TS; thus the AEWLIS sentence can be graphically represented by a matrix, having as many columns as the number of TSs and as many rows as the number of Tags.

¹For the sake of animation, unknown (out-of-signary) sign can be fingerspelled.

²This occurs for instance when the signer keeps the non-dominant hand, until a sign comes that requires both hands.

Italian	Per quanto riguarda i mari, generalmente mossi o molto mossi, poco mosso solo il Tirreno.										
AEWLIS	Position	1	2	3	4	5	6	7	8	9	10
	Lemma	MARE	PROPRIO	ZONA	IONIO	ADRIATICO	MOSSO	MOSSO	ZONA	TIRRENO	MOSSO
	Sign-ID	1349	1875	2100		3002	423	423	2100	3000	423
	Topic/Focus	topic		topic			focus	focus	topic		focus
	Facial							strong		mild	

Figure 2: AEWLIS annotation of a sentence “Concerning seas, generally from slight to moderate, smooth Tirrenian sea only”. Only the most significant and not empty Tags are reported. The sentence is practically split into three parts, Signs 1-2, 3-7 and 8-10. In each subparts there are specific topics and focuses. The general reference “generalmente” (“generally”) to the seas is interpreted by listing a few exemplars (Signs 4 and 5). The Facial Tag distinguishes between “mosso” and “poco/molto mosso” (empty, “mild”, and “strong”, Signs 6, 7 and 10, respectively).

6. Conclusion

This paper reported on the work by ATLAS project of defining a translation-oriented symbolic representation of LIS (AEWLIS) and of building a Italian-LIS parallel corpus annotated with AEWLIS. The AEWLIS language has been adopted both as an interchange format among translation and animation modules of the virtual interpreter and as annotation format for the corpus construction.

The first release of the corpus contains weather forecast bulletins, but ATLAS partners intend to significantly increase its size and extend it to other domains in order to make it a benchmark for further research on LIS. The corpus and other related linguistic resources developed as a side-effect of this research will be made freely available.

Acknowledgement

The work presented here has been developed within the ATLAS project, co-funded by Regione Piemonte within the “Converging Technologies - CIPE 2007” framework (Research Sector: Cognitive Science and ICT).

7. References

- BlueSign. 2010. <http://bluesign.dii.unisi.it/>.
- F. Brugnara, et al. 2000. Advances in automatic transcription of broadcast news. In *Proc. of ICSLP*, pp II:660–663, Beijing, China.
- BSLCorpus. 2010. <http://www.bslcorpusproject.org/>.
- J. Bungeroth, et al. 2006. A German Sign Language Corpus of the Domain Weather Report. In *Proc. of LREC*, pp 2000–2003, Genoa, Italy.
- Jan Bungeroth, et al. 2008. The ATIS Sign Language Corpus. In *Proc. of LREC*, Marrakech, Morocco.
- J. Chon, et al. 2009. Enabling access through real-time sign language communication over cell phones. In *43rd Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA.
- DictaSign. 2010. <http://www.dictasign.eu>.
- DizLis. 2010. <http://www.dizlis.it>.
- Echo. 2010. <http://www.let.kun.nl/sign-lang/echo>.
- ELAN. 2010. <http://www.lat-mpi.eu/tools/elan/>.
- eLIS. 2006. <http://elis.eurac.edu>.
- R. Elliott, et al. 2004. An Overview of the SiGML Notation and SiGMLSigning Software System. In *Proc. of LREC*, pp 98–104, Lisbon, Portugal.
- eSign. 2009. <http://www.sign-lang.uni-hamburg.de/esign/>.
- M. Filhol and A. Braffort. 2006. A sequential approach to lexical sign description. In *Proc. of the Workshop on Sign Languages, LREC*, Genoa, Italy.
- Eva Hajičová, et al. 1998. *Topic-focus articulation, tripartite structures, and semantic content*. Kluwer.
- HamNoSys. 2004. <http://www.sign-lang.uni-hamburg.de/projects/hamnosys.html>.
- Matt Huenerfauth. 2006. *Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation*. Ph.D. thesis, Computer and Information Science, University of Pennsylvania.
- P. Koehn, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL - Demo and Poster Sessions*, pp 177–180, Prague, Czech Republic.
- Leonardo Lesmo, et al. 2009. Legal modificatory provisions and thematic relations. In *ICON*, pp 352–357, Hyderabad, India.
- Diane Lillo-Martin and Ronice Müller de Quadros. 2004. Structure and acquisition of Focus in ASL and LSB. In *Proc. of TISLR*, Barcelona, Spain, October.
- Ian Marshall and Éva Sáfár. 2003. A prototype text to British Sign Language (BSL) translation system. In *Proc. of ACL*, pp 113–116, Morristown, NJ, USA.
- mobileASL. 2010. <http://mobileasl.cs.washington.edu>.
- Sara Morrissey. 2008. *Data-Driven Machine Translation for Sign Languages*. Ph.D. thesis, School of Computing, Dublin City University.
- NGTCorpus. 2010. <http://www.ru.nl/corpusngtuk/>.
- Volha Petukhova and Harry Bunt. 2008. Lyrics semantic role annotation: Design and evaluation of a set of data categories. In *Proc. of LREC*, Marrakech, Morocco.
- SignSpeak. 2010. <http://www.signspeak.eu/>.
- SignWriting. 2009. <http://signwriting.org>.
- William C. Stokoe, et al. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. Linstok Press., Silver Spring, MD, 2 edition.
- U.Boston. 2002. <http://www.bu.edu/asllrp>.
- U.DePaul. 2008. <http://asl.cs.depaul.edu/>.
- Tony Veale, et al. 1998. The Challenges of Cross-Modal Translation: English-to-Sign-Language Translation in the Zardoz System. *Machine Translation*, 13(1):81–106.

A Multilanguage Database for supporting Sign Language

Translation and Synthesis

**Roberto Borgotallo⁽²⁾, Carmen Marino⁽²⁾, Elio Piccolo⁽¹⁾,
Paolo Prinetto⁽¹⁾, Gabriele Tiotto⁽¹⁾, Mauro Rossini⁽²⁾**

(1) Politecnico di Torino, Department of Control and Computer Engineering
Corso Duca degli Abruzzi 24, Torino, Italy

(2) RAI Radiotelevisione Italiana, Centre for Research and Technological Innovation
Corso Giambone 68, Torino, Italy

r.borgotallo@rai.it, carmen.marino@rai.it, elio.piccolo@polito.it,
paolo.prinetto@polito.it, m.rossini@rai.it, gabriele.tiotto@polito.it

Abstract

The design of a language database is an important task within projects targeting sign language research. In this paper is presented a database structure that supports both linguistic information and visualisation oriented data to assist a final publication of services for deaf people. The database has been designed within the Automatic Translation into sign LAngeageS (ATLAS) project that takes aim at getting the automatic translation from written Italian to Italian Sign Language (LIS). The final step of the overall process is the enrichment of the original video with a superimposed virtual character realised by 3D animated computer graphics. The top element within the database is the A_Product defined as the main primitive element managed by the ATLAS platform under which all the other data, from input sources to the final publication modalities and attributes lay. The A_Product includes the reference to the original content and all the intermediate elaborations results towards the final publication comprehensive of the virtual character animations. Among the others, the most important transformation is the automatic translation from a written Italian text to the intermediate language AEWLIS (ATLAS Extended Written LIS), formalized within the ATLAS project.

1. Introduction

The automatic translation among national languages represents one of the greatest challenges undertaken by computer science. The automatic translation from Italian language into Italian Sign language, the mother tongue for signing deaf people, may be regarded as a venture even more difficult because the syntax, the grammatical structure and the lexical heritage of the two languages are very different. However, the request for Italian Sign Language (LIS) interpretation is increasing in different contexts nowadays, such as educational, legal, healthcare, entertainment and cultural environments.

In this paper is presented the structure of a Multilanguage database that supports the translation from Italian Language into Italian Sign Language. The adopted methodology and its structure can be extended to support the translation of each national language into the corresponding sign language. The database has been designed within the Automatic TransLation into italian Sign language (ATLAS) project that aims at provide the LIS translation of different typologies of contents, such as audio/video, subtitles, teletext pages, web pages, texts, and displays it through a virtual interpreter realized by 3D animated computer graphics. The ATLAS architecture allows the retention of linguistic information, including lexical and animation data into a unique database named Atlas MultiMedia Archive (AMMA).

The first section of the document is dedicated to the description of the basic process that transforms, through several intermediate steps, a generic source content into a

final virtual actor animation.

The second part concerns the structure of AMMA and contains the description of all included databases.

2. Language Translation Process

The translation process from Italian language into Italian Sign language is articulated into three main steps:

- Generic content ingestion
- Text Translation into AEWLIS
- Virtual Actor commands generation and rendering

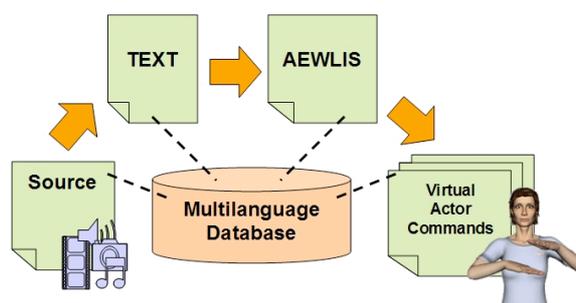


Figure 1: Translation chain

2.1 Generic Content Ingestion

The source content ingestion is the first phase of the translation chain that consists in the submission of different typologies of contents: audio, video, teletext pages, web pages, texts and subtitles. Whatever the format, the text is the fundamental part because it is

starting from it in order that the translation takes place. The multimedia components are as well important for the generation of an effective service over communication channels such as the digital television or mobile streams where the multimedia experience is mandatory.

In fact the ingested multimedia components just pass through the system when they are already suitable for the final service, otherwise they are adapted by mean of transcoding to satisfy the specific publication channel requirements (e.g. suppose that a video mpeg2 with standard definition is supplied, it is adequate for the DVBT channel but has to be transcoded to lower resolution and bit rates for mobile streaming).

As the text is a mandatory component of the ingestion phase, it is interesting to manage the creation of a text starting from the audiovisual content. Such an operation is possible with the adoption of an automatic speech recognition subsystem that derives the spoken words automatically from the audio signal. Usually after this stage, a manual revision is required because of the non-negligible error rate, both the automatic speech recognition and the manual revision are considered as *transformations*.

2.2 Translation into AEWLIS

Within the ATLAS project it has been formalized an intermediate language called AEWLIS that contains all necessary information to derive a good animation of the virtual interpreter.

The AEWLIS inherits the specific morphologic and syntactic structure of LIS and includes the so called Communication Channels that specify the position/direction of the hands, body, shoulders, head, gaze, labial and facial expressions that are very important in sign communications.

Translation from written Italian to AEWLIS text is the most tricky part of the overall process and it is based on very complex statistical and mathematical algorithms. There are basically two distinct approaches to fulfill this task: ruled based and statistical translation.

The first kind of translation is based on rules for mapping the grammar and syntactic structure of the input language to the output language (AEWLIS). Statistical translation is based instead on classical machine learning algorithms. In this case, after a preliminary learning phase, where a large number of manually translated phrases (corpora) are submitted, the machine translation system is expected to automatically generate the translation for new input sentences with a sufficient precision.

2.3 Virtual Interpreter Animation

Nowadays, through computer graphics, it is possible to model and animate a virtual character that reproduces the LIS movements. In addition to movements of fingers and hands, also arms and facial expressions can be reproduced in a detailed way. This is a crucial aspect because in LIS the mimic and gestural expressiveness of the body and face assumes an relevant importance. The

animations are automatically derived from the AEWLIS by mean of an articulated engine for the planning of the movements along the timeline, followed by the creation of virtual actor commands that are finally used for rendering.

3. Database structure and utilization

The database is an essential component to support the above mentioned translation process. It constitutes the persistence layer for all the sub processes implied, for example the manual annotation aimed at the creation and assessment of the corpora for the automatic statistic translation. Moreover, on the database are stored all the elaboration results like the Italian and AEWLIS texts, the commands for the virtual actor, the pointers to external references like Wordnet and Radutzky dictionary.

Figure 2 shows an overview of the database with a leading concept of A_Product under which all the other databases is are linked: Multimedia, Corpora, Radutzky, New Signs, Wordnet, AEWLIS, Animation.

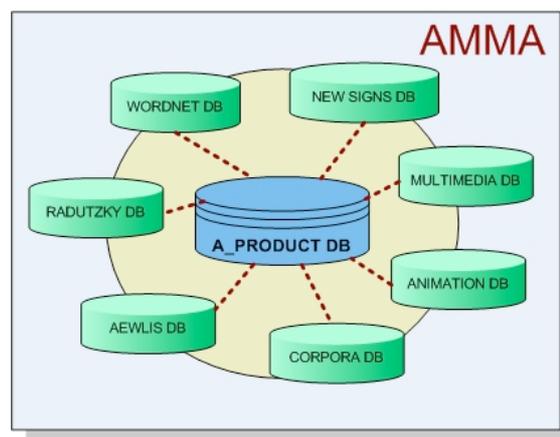


Figure 2:AMMA structure

Several other projects targeting sign language studies brought to the definition of a data storage and retrieval system, such as the Finnish Lexical Database (Savolainen & Leena, 2001), the Purdue American Sign Language Database for Sign Language recognition (Kak, 2002) and database for motion capture data retrieval (Award et al., 2009). These projects and know how constituted valid guidelines to organise the presented work.

3.1 A_Product Database

The A_Product is a data structure that contains all the information associated to the source content and its transformations, aimed at the generation of the virtual character animation. The A_Product includes the identification and process metadata that allow to keep trace of it during the different elaboration steps..

In figure 3 is shown in detail the structure of the A_Product.

It contains a reference to the multimedia archive and to the text resources, the provenence of these components is

tracked as well by mean of the source block in the figure 3. For example, in this way is possible to know that a text is coming from the Teletext of a certain channel and date or a multimedia is coming from a specific archive.

An A_Product instance is usually composed by a multimedia (e.g. a weather forecast edition) and a text representing what is said in the programme. In the above mentioned example the text could be derived automatically from the audio by an ASR (Automatic Speech Recognition) tool, in the picture this is represented by the “MM to TEXT TRANSF”. In other cases the A_product is formed only by text for example coming from Teletext or Web.

Both multimedia and text can be transformed respectively in order to get a suitable version of the multimedia for publication and to improve the automatic translation into LIS. This transformations are represented in the figure with a looping relation.

The main multimedia transformations are: audio track extraction, text to speech conversion, speech to text conversion, video transcoding and audio transcoding. As far as textual transformations are concerned, the most important are: text synchronization, subtitles adaptation , text manual revision and validation.

One of the most important step is the translation of the text into AEWLIS, better detailed in the paragraph 3.5. Starting from the AEWLIS text and the corresponding LIS Signs the animation process will produce the commands for the final virtual interpreter representation. The publication is the final step that allows to deliver a profitable service to the final user over different communication channels (DTV, Mobile, Web). A typical multimedia service is composed by a video with a superimposed virtual character.

It has been considered important to store all the elements and the results of the different transformations in order to be able to improve the translation and use this data in other projects and research activities.

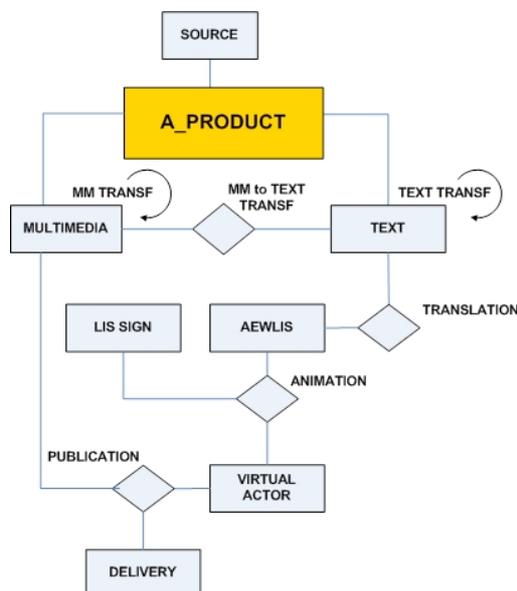


Figure 3:A_Product

3.2 Multimedia Database

This is the section where multimedia with different formats and resolutions are stored. Multimedia are saved on a common file system storage while they are referenced by pointers inside the database. The virtual actor animation usually does not belong to this section as it is rendered over the multimedia stream in real-time. Nevertheless in some cases the virtual actor could be superimposed when for example the device is not enough effective.

3.3 Corpora Databases

Here are preserved the annotations and manual translations that constitute the ground truth for the learning phase of the translation statistic engine. As the annotation is quite hard and long to carry out this is a very precious information.

Each enter of the Corpora Database contains:

- A pointer to a text file
- A pointer to the corresponding AEWLIS file
- A pointer to the associated movie with a LIS interpreter
- A pointer to the Virtual Character Commands file

A large set of metadata are collected resorting to the ECHO project regarding corpora definition metadata. For a deeper view on used metadata formats it has been considered the document “Metadata for sign language corpora” (Crasborn & Hanke, 2003).

3.4 New signs database

New signs are created if they are not present in the database. They are detected by means of a lexical frequency analysis. We linked the results of this analysis with the Radutzky Dictionary in order to find what signs are present within the domain in analysis, that are not present in the database yet. The new signs can be detected during the source texts analysis and during annotation. In this case a new sign has been agreed by the group of annotators including people of the deaf community. In the database is stored the new sign with its meaning and its video recording. The annotation tool used allows to easily detect the new signs by automatically searching into the database and retrieving the signs and comparing it with the signs being annotated. This tool called ALEA has been developed in the ATLAS project and represents a web based annotation editor.

To define a new sign is necessary to establish some parameters that indicate if it is:

- Size Modifiable
- Speed Modifiable
- Space Relocatable
- Compound Sign
- Symmetric Sign
- Static Sign
- Repeated Sign

Other additional information concerns the number of the involved hands, the movements of body, shoulder, head,

gaze and labial.

3.5 AEWLIS database

This is the section where the written LIS - the formal result of the language translation - is stored, subdivided into phrases. The visualization process starts from this formal representation with its conversion into virtual character commands. The stored data include the sentences written in AEWLIS and all the information that can be derived by the annotation in this formalism.

3.6 Radutzky and Wordnet databases

These are external databases, hooked to the system in order to increase the information. With Wordnet, each lemma is resolved with a better semantic representation while pointing to the Radutzky dictionary is possible to access to their basic sign representation. The Radutzky database contains the lemmas information resorting to the Radutzky Dictionary. Each sign is stored along with a coding that gives information about its parameters, such as:

- Hands Configuration
- Hands Orientation
- Place
- Movement

Each Radutzky sign is linked to the Wordnet Synset of each lemma to manage synonyms and perform disambiguation during manual translation. The association Radutzky sign-Wordnet synset allows to identify, for each new Italian word in the Italian source text, if it is present or not in the standard LIS dictionary. If it is absent, we can automatically find Italian synonyms that have a correspondence in the Radutzky dictionary. This process facilitates new sign creation and the use of other signs that are not standard but widely used in the deaf community.

3.7 Metadata

The database allows to store all the metadata related to the products created within the process. The metadata are always associated with a transformation (MM to text, textual transformations, etc...). They are used to store the information concerning the output of a specific transformation. A set of metadata is defined in order to store information on the product creation, date and time.

4. Conclusion

The presented database allows to share in a structured way information regarding the automatic translation process from Italian to LIS. The design of the database allowed to deeply investigate relations and dependencies between the major entities taking part into the translation process.

It supports all the operations from the content ingestion to the visualization by storing all the necessary data and supporting the transformations that are needed in the translation process. It provides the storage of metadata associated to each transformation that are useful to trace

information about the creation of each product.

The modular structure of the proposed database and related processes allows to extend all the elaborations to other natural languages and other sign language dialects. We assume this is feasible with straightforward modifications and inclusion of additional databases connected to the A_Product.

The database data storage phase is still ongoing and future work will aim at the definition of a procedure in order to store critical data (i.e. new signs with associated movie and meaning) during manual processes, such as annotation and source text analysis.

5. Acknowledgements

The work presented in the present paper has been developed within the ATLAS (Automatic Translation into sign LAnguageS) Project, co-funded by Regione Piemonte within the "Converging Technologies - CIPE 2007" framework (Research Sector : Cognitive Science and ICT).

6. References

- Savolainen, Leena, (2001). The database system used in the Finnish Sign Language Dictionary Project. *Sign Language and Linguistics*
- Kak, (2002), 4th IEEE international Conference on Multimodal interfaces. *Purdue RVL-SLLL ASL Database for Automatic Recognition of American Sign Language*.
<http://dx.doi.org/10.1109/ICMI.2002.1166987>
- Awad, C., Courty, N., Duarte, K., Le Naour, T., Gibet, S.(2009). *A Combinated Semantic and Motion Capture database for Real-Time Sign Language Synthesis*. Université de Bretagne Sud, Vannes, France.
- Crasborn, Hanke (2003). Background document for an ECHO workshop. *Metadata for sign language corpora*, Radboud University Nijmegen
http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf

A Corpus for Verifying American Sign Language During Game Play by Deaf Children

Helene Brashear¹, Zahoor Zafrulla¹, Thad Starner¹,
Harley Hamilton¹, Peter Presti² and Seungyon Lee¹

¹Georgia Institute of Technology, College of Computing, GVU, Atlanta, Georgia, USA.
(brashear, zahoor, thad, harley.hamilton)@cc.gatech.edu

²Georgia Institute of Technology, Interactive Media Technology Center, Atlanta, Georgia, USA.
peter.presti@imtc.gatech.edu

Abstract

The CopyCat project was designed to develop an interactive educational adventure game to help deaf children acquire language skills. The main goals of the project are to improve the language and memory abilities of deaf signing children, advance basic research in computer-based sign language recognition, and design an efficient language interaction model in order to assist in the language learning of deaf children. The CopyCat project was begun as a collaboration between Georgia Tech and the Atlanta Area School for the Deaf in 2004 and has been collecting ASL (American Sign Language) data since Spring of 2005. Since then we have collected 5829 signed phrases from over 30 children. In this paper we describe the evolution of the CopyCat system design, data collection methodology, and resulting corpus, as well as challenges and successes throughout the process.

1. Introduction

It is important that children are exposed to sufficient language examples during early childhood to aid in the development of life long language skills. Language learning is dependent upon the availability of that language and the opportunities a child (Spencer and Lederberg, 1997) or an adult learner (Krashen, 1980) have for interacting with skilled users of the language. This “critical period” of language exposure is important for both spoken and signed languages (Mayberry and Eichen, 1991; Newport, 1990). Ninety percent of deaf children are born to hearing parents who may not know sign language or have low levels of proficiency with sign language (Gallaudet, 2001). Many of these deaf children of hearing parents remain significantly delayed in language development due to a lack of language exposure at home. For many of these children the first consistent exposure to quality language models will be when they enter school, which can result in lifelong difficulties with communication (Stinson and Foster, 2000).

CopyCat was designed to address these language learning issues by facilitating the development of both expressive language and working memory skills. While computer-child interaction cannot replace high quality adult-child interaction, it can be designed to integrate meaningful authentic communication in order to enhance expressive language and working memory skills that may facilitate the child's ability to make the most of opportunities for acquiring language in a natural way via human interaction.

2. Evolution of CopyCat System

Our ASL data is collected on-site at schools around the Atlanta area. Children play a computer game by wearing colored gloves and signing to characters within the game to accomplish game objectives such as rescuing kittens or defeating villains such as alligators and snakes. Data is collected via wireless accelerometers mounted on the wrists of the gloves and a single video camera. The sensor data is collated and time stamped by the game system and saved

as our library for developing our recognition system and linguistic review.

The system has been built in three main design phases: each phase addresses game design, data collection, and the ASL recognition engine. Each iteration has been designed with the ultimate goal of moving towards a fully functional system with live recognition that provides productive feedback for students of varying skill levels.

Our corpus collection methods were designed to attempt to elicit live, natural signing from children as they interact with characters in the game. This approach has resulted in a data set that contains many language modeling challenges including disfluencies, pauses, dominant hand switching, and sign variations. Our research has focused on developing labeling schemes and training models to accurately reflect the children's signing. A prototype live recognizer developed from this corpus has been deployed for testing and has been shown to have a statistically significant educational effect on language learning as compared to a control condition.

2.1. The CopyCat Games

As part of the CopyCat project, several computer-assisted language learning games have been designed. Each game entails some sort of quest by the hero to collect items in order to remediate a problem. In each quest, the children interact with the hero via sign language to tell the hero warn them of a villain or identify where a hidden object is located. If the children know what to tell the hero regarding the guards location they can push a “talk button to turn the hero towards them so they can sign to him/her. They then push the “talk button again when they are finished signing. If the children are uncertain what to say they can click a “help button to see the tutor in the top left corner of the screen tell them what to say. The child may view the tutor repeatedly if (s)he so chooses (see Figure 1).

After the child talks to the hero, the child's signing is classified as correct or incorrect. If the child's utterance is in-

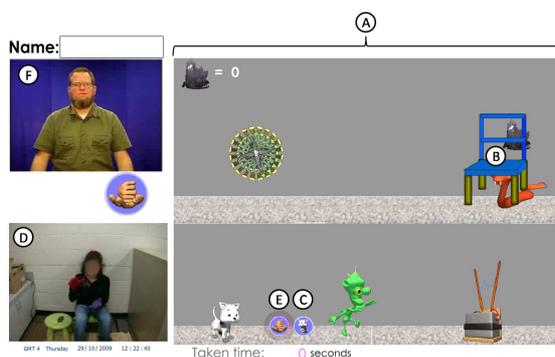


Figure 1: CopyCat screen shot from Mini Quests: A) Animated game characters in their worlds B) The villain (a snake) is hiding under the chair. C) The push to sign button has a picture of the kitten on it. Children push the button to sign to the kitten and warn her about the snake. D) The live video feed allows children to see themselves as they sign. E) The help button has a picture of the sign for help and will bring up ASL video to help the children during game play. F) This window is the help video feed.

correct, a question mark appears above the hero's head, to simulate misunderstanding by the hero, and the child must try again to communicate accurately. If the child's sign is correct, the hero, with the wave of a paw, "poofs" the guard, turning it into an innocuous item and the hero continues on the quest

2.2. Language Learning

The video tutor examples in the game were designed to be similar to a communication setting which young children encounter while learning language through interaction with adults. As the child's linguistic and communicative competence and confidence grow, the need for such assistance diminishes and the child can respond appropriately without help. Thus, our tutor performs the role of the good adult language model (Schieffbusch and Bricker, 1981), always available to the child, responding to the child's cue (in this case a press of the "help" button) in an appropriate linguistic manner.

2.3. Educational Evaluation

In order to collect data regarding the language processing abilities of the children and the efficacy of the game's language interaction model, pretests and post tests were administered and in-game response data were recorded. These tests consisted of sections to test receptive language skills, expressive language skills, and working memory. The results of the expressive language test indicate that the experimental group made a significant gain in the accuracy of their utterance to describe the video they saw as well as in their length of utterance as measured by mean length of utterance from pretest to post test (Weaver et al., 2010).

3. System Design

3.1. Iterative Design Cycle

The iterative design cycle allows us to adapt to problems as they emerge during the development process and has allowed the CopyCat system to improve rapidly.

3.2. Interface Design

Our user interface for game play uses a video stream of the user, feedback from characters in the game, and help videos in ASL to engage the children. The live video stream allows the children to see their signing and engages them in the signing. The children enjoy "being in the game" and tend to use the feedback to stay in frame.

The game characters have been designed to attempt to elicit natural signing. When the child pushes the signing button, the character will face the child and pay attention while (s)he is signing. If the signing is incorrect, a question mark thought bubble shows above the character. We have found that visual clues such as these help guide the children in their interactions.

The introduction instruction and game help videos are all ASL. We have taken care to synchronize the spatial layout of the game with the spatial constructs in signing to provide consistency. Even simple modifications to the interface such as moving a button require a check of all of the ASL spatial referencing in the videos.

3.3. Wizard of Oz

When the functionality of a system is under development, developers can sometimes replace that functionality with a person, similar to the "Great Wizard of Oz" operating behind the curtain. The system can be tested while the hidden "wizard" controls operations and developers can obtain critical feedback about system design early in the process (Dix et al., 2004).

We divided game development and sign language recognition by using a "Wizard of Oz" setup, shown in Figure 2 (Henderson et al., 2005). The child interacts with the user computer (on the right) by navigating with the mouse and signing to characters. The wizard's computer (on the left) controls the game's response to children signing and collects data from the sensors and game logs for future use.

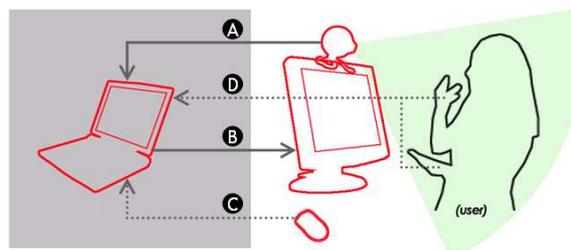


Figure 2: Diagram of the Wizard of Oz system setup showing a) live camera and sensor feed b) interface output split between wizard and user c) child's mouse and d) the interface computer

3.4. Sensors

The CopyCat system uses computer vision and three-axis accelerometers to collect data for use in sign language recognition. Our computer vision is processed from video collected on a single camcorder that faces the children. The children wear colored gloves, which contain small accelerometers mounted on the outside of the wrist (shown in Figure 3). These accelerometers provides information on movement acceleration, direction, and rotation of the hand. The distinct color of the gloves helps distinguish the hands from the skin color of the face and cluttered backgrounds. The wizard’s computer coordinates the data streams, synchronizes them, and stores them for future use.



Figure 3: Gloves with accelerometer (top). Close up of wrist-mounted accelerometers (bottom).

One key design goal has been to have a portable system that will work in a variety of environments. Our deployment environments at the schools have ranged from classrooms and libraries, to a re-purposed supply closet. Figure 4 shows the construction of a “signing kiosk” and the resulting view from the camera. The kiosk is inexpensive and modular so that it can be transported easily. The kiosk fixes the position of the camera relative to the child’s position on the chair. The color of the furniture can be used to help calibrate the video camera’s color balance to enable better hand tracking. This kiosk design allows us to move our equipment from area to area with minimal re-calibration.

4. Resulting Corpus

4.1. Overview of data collected

Each phase of the CopyCat project includes on-site deployments to collect data at our two partner schools. We have



Figure 4: Kiosk setup

Subject	Object	Adjective	Verb
alligator	bed	black	behind
cat	box	blue	in
snake	chair	green	on
spider	flowers	orange	under
	wagon	white	
	wall		

Table 2: Game vocabulary

collected a total of 5829 phrases over four phases, with a total of nine deployments. Table 1 shows a count of phrases collected throughout the CopyCat project. Each phrase is a three, four, or five sign sentence taken from a vocabulary of 22 signs. The phrases are of the format *[adjective1] subject preposition [adjective2] object*.

4.2. Characterizing the Children’s Signing

Most of the sign language databases used for automatic sign recognition are carefully scripted and collected in a controlled environment (Holt et al., 2006). Our data set provides many samples of children signing as they interact with the online characters. This signing contains many of the artifacts of conversational signing such as disfluencies like pauses, false starts, hesitations, and sign variations. It also has many examples of non-signing activities such as scratching and fidgeting.

The conversational nature of the children’s interaction with characters results in signing samples that contain signing beyond basic game vocabulary. The data set contains many non-game communications towards game characters (including messages such as “wrong”, “start again” or “not red I mean blue”), and even gestures that are not ASL such as a wave which is used generally to indicate an error and restart (a kind of “erase” gestures).

The children’s signing handedness did not directly correspond to their dominant handedness for other activities and was inconsistent even within the phrases. This hand switching makes it more difficult to group signs and phrases by handedness for modeling purposes. Dominant hand switching is probably a symptom of their low fluency and is common among children (Mandal et al., 1999).

Phase	Game	Date	Participants	Ages	Total Phrases
Pilot	Kitten Escape!	Spring 2005	3	9-11	50
Pilot	Kitten Escape!	Spring 2005	2	9-11	78
Total					128
First Deployment	Kitten Escape!	Spring 2005	5	9-11	627
First Deployment	Castle Quest	Fall 2005	9	9-11	1812
Total					2439
Second Deployment	MiniQuests	Fall 2008	5	6-9	505
Second Deployment	MiniQuests	Spring 2009	5	6-9	503
Second Deployment	MiniQuests	Spring 2009	14	6-9	822
Total					1830
Third Deployment	MiniQuests	Fall 2009	11	6-9	1432
Total					1432
CopyCat Total					5829

Table 1: Table of data collected during the CopyCat project

4.3. Annotation

Our current annotation system is designed for sign classification, recognition, and verification. First we label each sign by its English label (green, cat, etc.). The initial label set has been expanded to include non-game vocabulary from children, as well as some non-ASL gestures such as pauses, fidgets, and waves. Signs are then annotated for handedness by hands used during the sign and the dominant hand: right hand, left hand, both+right hand dominant, both+left handed dominant, both+symmetric. Finally signs are rated for quality as good, ok, or bad.

5. Using the Data

5.1. ASL recognition

Our first task with the data set is automatic sign language recognition. In this process, we collect samples of signs, train up models using the samples, and then use the models for recognition. When the models are trained we use an independent test set for validation results. This means that we divide the data set into one group for training the models and another group for testing the models in order to see how well the models perform against signs examples that are previously unseen to the computer (Brashear et al., 2006).

5.2. ASL verification

Our second task is automatic sign language verification. In this process, we collect samples of signs, train up models using the samples, and then use the models to verify sign samples as a correct match or incorrect match to a baseline phrase. To get the verification we run data for a sample against the expected model and the use a common rejection threshold on the likelihood.

5.3. Tests of live system

In Fall of 2009 we conducted our first pilot tests of the live system. The verification system was based on models built

with data from our first deployment. The results of that test are currently being compiled.

6. Challenges of the CopyCat Corpus

6.1. Library Continuity

There is a continued tension between goals for system improvement, expansion of game functionality, and library expansion. Though our upgrades in sensors and configuration have improved the reliability and portability of the system, they also detract from backwards compatibility. This discontinuity results in a larger corpora of children signing, with sub-sets from various deployments that are incompatible with each other.

The library data is stored in both its raw format as well as a format that includes post-processing from vision and accelerometer sub-routines. This redundancy in storage requires more disk space, but helps alleviate the continuity problems by allowing for changes in post-processing without losing entire library sets. For example, we have changed our computer vision code several times. The raw data library allows us to experiment different post-processing schemes and choose optimally.

6.2. Sensor Changes

During the design cycle we have changed the sensors several times. Two of our main design priorities are system reliability and system portability. Our long term goal is a system that can be set up at any school and requires minimal maintenance. We started with the explicit goal that our sensors be inexpensive and easy for schools to use.

During the project, we have used both commercially available accelerometer and those we design in-house. We have gone through several iterations of accelerometer collection code in order to address issues that emerged with calibration, output normalization, and sensor drift (Westeyn et al., 2009). These changes, combined with changes to the video frame rate, create incompatibilities with existing data from

previous deployments. This is a further challenge to library continuity and such changes must be carefully considered.

6.3. Varied Environments

Each time we visit a school for a deployment, we have no guarantees where they will have space for us to set up. These changes in environment create challenges for computer vision algorithms. Many sign language recognition systems depend on very static environments for their algorithms to work. We have worked to make the system more portable by a combination of choosing more flexible algorithms and creating an environment where visual cues can help keep the algorithms calibrated. The kiosk helps ensure that the camera distances are approximately the same each time. Additionally the colored gloves and furniture help provide reference points for algorithms to track hands, face, and body movement in the video frame.

6.4. Data Integrity

During data collection the system must coordinate data streams from three different sensors. These streams must be saved to disc, logged, and synchronized. One of the challenges of this configuration is keeping the data streams synchronized and providing live feedback for errors in reading, synchronizing, or logging sensor data. Our most recent iteration has focused on creating a subsystem specifically to provide feedback to administrators to prevent problems in game play and data loss.

Post-processing of the libraries can also discover errors in the data stream. These errors must be diagnosed for future prevention and the samples must be catalogued as damaged data.

6.5. Automatic Annotation

We have designed the game to provide as much automatic annotation as possible to help us index and use our data. Each signed phrase contains logs with information on user, session details, wizard feedback, and game information. After the data is stored, our post-processing is also largely automated. These logs provide further information about the content of the signed phrase. All of this data helps us rapidly compile statistics on the data set and pull out subsets by interesting features.

6.6. Maintaining library

As the library increases in size and complexity we have continued to try to address issues with maintaining our data. Maintaining logs and raw data allow us to continue to do retrospective evaluations of many aspects of the process. There are different research and publication cycles for the various topics of the CopyCat project: computer vision, machine learning, human-computer interaction, sign linguistics, and education.

The size of the data has been growing since the beginning of the project. Not only does each deployment add more data instances to the library, but the size of the data per instance has been growing as well. Verifying the integrity of automated process logs is tedious and is time consuming. We have increased sensor sampling rate, as well as the detail and complexity of the game logs. Additionally, we must

keep track of data from educational testing which includes a large amount of video of the children's language testing sessions.

6.7. Sign Variation

The machine learning system needs many examples of the same signs across many systems for building representative models that are robust to variations. Thus far we have maintained a fairly small vocabulary, which allows for many examples of a sign. Even with the small vocabulary, we have discovered that there are often many variations on how a sign is performed. Most of these variations are technically correct and we must make allowance for them. If only one or two children perform a specific variation, it can make collecting sufficient examples difficult.

6.8. Developing Annotation Schemes

One of the goals of the machine learning research is develop generalized annotation schemes that will scale with larger data sets and vocabulary. Experimenting on this front can be very challenging since annotation schemes aren't standardized and the conversational nature of the children's signing creates unexpected variations in sign structure and performance. Annotating large sets of data can be time consuming and tedious. We have created an in-house annotation tool that acts like a video editor and can add multiple tags to the same sign sequence to indicate various labels such as the sign name, handedness, and quality. This tool allows for the addition of new tags as the annotation scheme evolves. Additionally, the collection of tags can be used to create different model groups for classification. For example a time sequence could be modeled as "cat", "cat" with both hands, or "cat" of good quality. We can test these variations in modeling to compare their performance.

6.9. Influencing the Children's Signing

Throughout the iterations of game design, we have continued to create an interface that influences how the children sign. The story line of the game helps restrict vocabulary by limiting the scope of objects and characters on the screen for the children to describe or address. By creating a conversational environment, we can influence how the children sign. The "click to sign" approach to the game provides a dual purpose of segmenting the signing sequences and giving the children pause to focus. We have even found that children will sometimes take a moment to rehearse their signing before clicking to get the character's attention in the game. These techniques have greatly improved the quality and kind of signing we get from the children, but we still face challenges with out of vocabulary signs and the children's difficulties performing the signs correctly.

6.10. Live Testing

As the machine learning research progresses, we will begin to conduct more live tests of the recognition system. We have recently augmented our system so that we can collect data while the live tests are being conducted. This multi-tasking allows us to continue to catalogue data while we test our machine learning system.

6.11. Privacy Issues

Because our data is collected from children our data is subject to strict privacy requirements. Our long term goal is to make sections of the data available to linguistic and machine learning researchers. Anonymizing the video data compromises the content, since the face is the center of the signing space and facial gestures are a component in ASL. We have been working with our institutional review board and the host schools to create an agreement that would allow us a mechanism to release data to other researchers.

7. Conclusion

CopyCat is a long-term project that has used an iterative development to design an interactive, educational game for deaf children. Designing and deploying the game for user testing has created unique challenges in collecting, storing, and using the large data set of children's signs. We have addressed many of these challenges with strategic game improvements generated from the feedback phase of the iterative cycle.

As CopyCat matures into a commercial-grade system, we are focusing on long-term library collection and management. The success of CopyCat will depend on our ability to easily integrate new data from each deployment into our library. We are focusing on ways to automate the collection and indexing of data for storage in a central library. As we build models off of the central library, each deployment site will get updates to the game recognition system.

8. Future Work

We are currently reviewing data collected from the most recent deployment as well as the results of the live system tests. Our long term goals include expanding the number of students and creating new games. We are working to expanding the vocabulary and language structure in new games. Additionally we will be performing more user testing on the live recognition system to determine its educational efficacy and to further examine the user experience when the Wizard is removed from the loop.

9. Acknowledgements

This work is supported by the NSF, Grants #0093291, #0511900, and the RERC on Mobile Wireless Technologies for Persons with Disabilities, which is funded by the NIDRR of the US Dept. of Education (DoE), Grant #H133E010804. The opinions and conclusions of this publication are those of the grantee and do not necessarily reflect those of the NSF or US DoE.

Special thanks to students and staff at the Atlanta Area School for the Deaf, Georgia School for the Deaf, and Gwinnett ISD for their generous help with this project.

10. References

- Brashear, H., Park, K.-H., Lee, S., Henderson, V., Hamilton, H., and Starner, T. (2006). American Sign Language Recognition in Game Development for Deaf Children. In *Assets '06: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 79–86, New York, NY, USA. ACM Press.
- Dix, A., Finlay, J., Abowd, G., and Beale, R. (2004). *Human-Computer Interaction*, chapter 6.4 Iterative Design and Prototyping. Prentice Hall.
- Gallaudet (2001). Gallaudet University. Regional and National Summary Report of Data from the 1999–2000 Annual Survey of Deaf and Hard of Hearing Children and Youth. Washington, D. C.
- Henderson, V., Lee, S., Brashear, H., Hamilton, H., Starner, T., and Hamilton, S. (2005). Development of an American Sign Language Game for Deaf Children. In *IDC '05: Proceeding of the 2005 Conference on Interaction Design and Children*, New York, NY, USA. ACM Press.
- Holt, G. T., Hendriks, P., and Andringa, T. (2006). Why Don't You See What I Mean? Prospects and Limitations of Current Automatic Sign Recognition Research. *Sign Language Studies*, 6(4).
- Krashen, S. (1980). The Theoretical and Practical Relevance of Simple Codes in Second Language Acquisition. In Scarcella, R. and Krashen, S., editors, *Research in Second Language Acquisition: Selected Papers of the Los Angeles Second Language Acquisition Research Forum. Issues In Second Language Research*. Newberry House, Rowley, MA.
- Mandal, M. K., Asthana, H. S., Dwivedi, C. B., and Bryden, M. P. (1999). Hand Preference in the Deaf. In *Journal of Developmental and Physical Disabilities*, volume 11.
- Mayberry, R. I. and Eichen, E. B. (1991). The Long-Lasting Advantage of Learning Sign Language in Childhood: Another Look at the Critical Period for Language Acquisition. *Journal of Memory and Language*, 30:486–498.
- Newport, E. L. (1990). Maturation Constraints on Language Learning. *Cognitive Science*, 14:11–28.
- Schiefelbusch, R. L. and Bricker, D. D. (1981). *Early Language: Acquisition and Intervention*. University Park Press, Baltimore.
- Spencer, P. and Lederberg, A. (1997). Different Modes, Different Models: Communication and Language of Young Deaf Children and Their Mothers. In Romski, M., editor, *Communication and Language: Discoveries from Atypical Development*, pages 203–230. Harvard University Press.
- Stinson, M. and Foster, S. (2000). Socialization of Deaf Children and Youth in School. In Spencer, P. and Marschark, M., editors, *The Deaf Child in the Family and at School: Essays in Honor of Kathryn P. Meadow-Orlans*, pages 151–174. Lea Lawrence Erlbaum Associates, London.
- Weaver, K. A., Hamilton, H., Zafrulla, Z., Brashear, H., Starner, T., Presti, P., and Bruckman, A. (2010). Improving the Language Ability of Deaf Signing Children through an Interactive American Sign Language-Based Video Game. In *Proceedings of 9th International Conference of the Learning Sciences*.
- Westeyn, T., Presti, P., and Starner, T. (2009). A Naive Technique for Correcting Time-Series Data for Recognition Applications. In *Proceedings of the Thirteenth IEEE International Symposium on Wearable Computers (ISWC 2009)*. IEEE Computer Society.

Employing signed TV broadcasts for automated learning of British Sign Language

Patrick Buehler¹, Mark Everingham², Andrew Zisserman¹

¹Department of Engineering Science, University of Oxford, UK

²School of Computing, University of Leeds, UK

patrick@robots.ox.ac.uk

Abstract

We present several contributions towards automatic recognition of BSL signs from continuous signing video sequences: (i) automatic detection and tracking of the hands using a generative model of the image; (ii) automatic learning of signs from TV broadcasts of single signers, using only the supervisory information available from subtitles; (iii) discriminative signer-independent sign recognition using automatically extracted training data from a single signer. Our source material consists of many hours of video with continuous signing and aligned subtitles recorded from BBC digital television. This is very challenging material *visually* in detecting and tracking the signer for a number of reasons, including self-occlusions, self-shadowing, motion blur, and in particular the changing background; it is also a challenging *learning* situation since the supervision provided by the subtitles is both weak and noisy.

1 Introduction

The goal of this work is to automatically learn British Sign Language (BSL) signs from TV footage using the supervisory information available from subtitles broadcast simultaneously with the signing (see Figure 1). Previous research in sign language recognition has typically required manual training data to be generated for the sign *e.g.* a signer performing each sign in controlled conditions – a time-consuming and expensive procedure.

The main idea is to use a given English word to select a set of subtitles which contain the word – these form the positive training set – and a much larger set of subtitles that do not contain the word – these form the negative set. The sign that corresponds to the English word is then found using a multiple instance learning approach. This is a tremendously challenging learning task given that the signing is continuous and there is certainly not a one to one mapping between signs and subtitle words.

In order to learn a sign we require that it is signed several (more than 5) times by a single signer within one broadcast. However, we show that by adding an additional discriminative training phase, we are able to recognize this sign when signed by new signers within a restricted temporal search region.

Previous work on automatic sign extraction has considered the problem of aligning an American Sign Language sign with an English text subtitle, but under much stronger supervisory conditions (Farhadi and Forsyth, 2006; Nayak et al., 2009). Cooper and Bowden (2009) aim to automatically learn signs using the a-priori data mining algorithm, although without hand shape cues.

Outline. Knowledge of the hand position and hand shape is a pre-requisite for automatic sign language recognition. Section 2 presents our method for hand detection and tracking which uses a generative model of the image, accounting for the positions and self-occlusions of the arms. The results using this method exceed the state-of-the-art for the length and stability of continuous limb tracking.



Figure 1: **Example results.** The signs for “golf” and “tree” performed by two different signers are learned automatically. Our data is TV footage with simultaneously broadcast subtitles. Using an upper body pose estimator (Section 2), we find the location of the hands and arms in all frames. Knowing the hand position in each frame, signs are automatically learned from TV footage using the supervisory information available from subtitles (Section 3). With this method, a large number of signing examples can be extracted automatically, and used to learn discriminative sign classifiers (Section 4).

Section 3 describes our method for learning the translation of English *words* to British Sign Language *signs* from many hours of video with simultaneous signing and subtitles (recorded from BBC digital television). A multiple instance learning framework is used to cope with the misalignment between subtitles and signing and noisy supervision. Using the method we can learn over 100 signs completely automatically.

Lastly, Section 4 shows how the automatic recognition of signs can be extended to multiple signers. Using automatically extracted examples from a single signer we train discriminative classifiers and show that these can successfully recognize signs for unseen signers.

2 Hand and arm detection

In this section we describe our method for locating a signer’s hands in the video. Previous approaches to hand tracking have applied skin colour models (Cooper and Bowden, 2007; Holden et al., 2005; Farhadi et al., 2007;

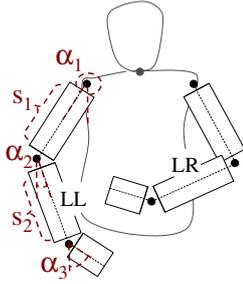


Figure 2: **Upper body model.** The pose is specified by 11 parameters – 5 for each arm and an additional binary parameter b indicating which arm is closer to the camera and hence visible in the case that the arms overlap. The shape of the head and torso and position of the shoulders are estimated in a pre-processing stage separate to estimation of the arm configuration.

Starnier et al., 1998) or sliding window hand detectors (Kadir et al., 2004). These methods perform poorly when the hands overlap or are in front of the head, and lose track due to the ambiguities that routinely arise, resulting in poor estimation of hand position or unreliable assignment of hands to left or right. In contrast, by using a full upper body model (Figure 2) and accounting for self-occlusion our method proves capable of robust tracking for long videos, *e.g.* an hour, despite the complex and continuously changing background (the signer is overlaid on the TV programme). Figure 5 shows example output of the tracker.

The remainder of this section outlines our upper body pose estimator which tracks the head, torso, arms and hands of the signer; further details can be found in Buehler et al. (2008). In the following, we refer to the arm on the left side of the image as the “left” arm, and respectively the arm on the right side of the image as the “right” arm.

2.1 Approach

Estimation of the signer’s pose is cast as inference in a graphical model of the upper body. To reduce the complexity of modelling and inference, the pose estimation process is divided into two stages (see Figure 3): (i) the shape of the head and torso and the position of the shoulders are estimated using a 2-part pictorial structure. This is relatively straightforward, and is described in Buehler et al. (2008); subsequently, (ii) the configuration of both arms and hands are estimated as those with maximum probability given the head and torso segmentations.

Generative model. Formally, given a rectangular sub-image \mathbf{I} that contains the upper body of the person and background, we want to find the arm and hand configuration $\mathbf{L} = (b, \mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_n)$ which best explains the image, where $\{\mathbf{l}_i\}$ specifies the parts (limbs) and b is a binary variable indicating the depth ordering of the two arms. In our application we deal with $n = 6$ parts: the left and right upper arms, the lower arms and the hands. The appearance (*e.g.* colour) and shape of the parts are learned from manual annotation of a small number of training images. The background is continuously varying, and largely unknown.

Every part $\mathbf{l}_i = (s_i, \alpha_i)$ is specified by two parameters: scale (*i.e.* length of a part modelling foreshortening) s_i and rotation α_i , and by the part to which it is connected. The connections are in the form of a kinematic chain for the left and right arm respectively (see Figure 2).

We define the probability of a given configuration \mathbf{L} conditioned on the image \mathbf{I} to be

$$p(\mathbf{L}|\mathbf{I}) \propto p(\mathbf{L}) \prod_{i=1}^N p(\mathbf{c}_i|\lambda_i) \prod_{j \in \{LL, LR\}} p(\mathbf{h}_j|\mathbf{l}_j) \quad (1)$$

where N is the number of pixels in the input image, \mathbf{c}_i is the colour of pixel i , and \mathbf{h}_j is a HOG descriptor computed for limb j (see below).

The formulation incorporates two appearance terms (described in more detail below) modelling the agreement between the image \mathbf{I} and configuration \mathbf{L} . The first, $p(\mathbf{c}_i|\lambda_i)$, models the likelihood of the observed pixel colours. Given the configuration \mathbf{L} , every pixel of the image is assigned a label $\lambda_i = \Lambda(\mathbf{L}, b, i)$ which selects which part of the model is to explain that pixel (background, torso, arm, etc.). The depth ordering of the two arms is given by the binary variable b which specifies which arm is closer to the camera and hence visible in the case that the arms overlap. The “labelling” function $\Lambda(\mathbf{L}, b, i)$ is defined algorithmically essentially by rendering the model (Figure 2) in back-to-front depth order (the “painter’s algorithm”) such that occlusions are handled correctly. For a given pixel, the colour likelihood is defined according to the corresponding label. Note that the pixel-wise appearance term in Eqn. 1 is defined over *all* pixels of the image, including background pixels not lying under any part of the model.

The second appearance term, $p(\mathbf{h}_j|\mathbf{l}_j)$, models the likelihood of observed gradients in the image (Figure 3c). This is based on Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) templates for the left and right lower arms, learned individually for different angles and scales. The HOG descriptor captures local information about image edges and shading with a controlled degree of photometric and spatial invariance. By using these descriptors, we exploit both boundary and internal features to determine the position and configuration of a limb.

The third term, $p(\mathbf{L})$, models the prior probability of configuration \mathbf{L} . This places plausible limits on the joint angles of the hands relative to the lower arms, and enforces the kinematic chain.

Complexity of inference. There are 11 degrees of freedom in the model: 5 for each arm and 1 for the depth ordering. The state spaces of the arm parts are discretised into 12 scales and 36 orientations. The hand orientation is restricted to be within 50 degrees relative to the lower arm and discretised into 11 orientations. Hence, the total number of possible arm configurations is $2 \times ((12 \times 36)^2 \times 11)^2 \approx 10^{13}$. Brute force optimisation over such a large parameter space is not feasible – the method described in the next section addresses this problem.

2.2 Computationally Efficient Model Fitting

The vast number of possible limb configurations makes exhaustive search for a global minimum of the complete

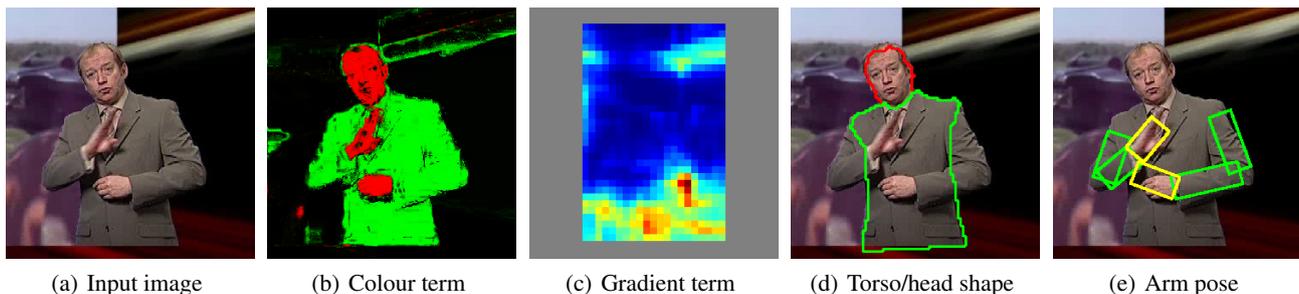


Figure 3: **Overview of pose estimation process.** Pose estimation for a given image (a) is performed using colour-based likelihoods (b) and likelihoods based on image gradients (c). The colour term in (b) is visualised by assigning the posterior probability for skin and torso to red and green colour channels respectively. The visualisation of the gradient term in (c) shows, for a given HOG template with fixed orientation and foreshortening, the likelihood at all locations in the image, where red indicates high likelihood. The example shown is for the right lower arm with rotation and foreshortening set to the ground truth values. Note the maximum is at the true centre point of the right lower arm in the image. Using the colour term (b) the head and torso can be segmented (d). The arm pose (e) is then estimated using the estimated torso and head shape, and both colour and gradient terms.

cost function infeasible. We therefore propose a fast approach based on a *stochastic* search for each arm, using an efficient sampling method (Felzenszwalb and Huttenlocher, 2005) to propose likely candidate configurations. Tree-structured pictorial structures are well suited for this task since samples can be drawn efficiently from this distribution (Felzenszwalb and Huttenlocher, 2005). However, they have several shortcomings explained in Buehler et al. (2008), e.g. the over-counting of image evidence. We show that by *combining* a sampling framework to hypothesise configurations with our full modelling of occlusion and background to assess the quality of the sampled configurations, we obtain the robustness of our complete generative model with the computational efficiency of tree-structured pictorial structure models.

The posterior distribution from which samples are drawn is given in Felzenszwalb and Huttenlocher (2005) as

$$p(\mathbf{L}|\mathbf{I}) \propto p(\mathbf{L}) \prod_{i=1}^n p(\mathbf{C}_i|\mathbf{l}_i) \quad (2)$$

where $\mathbf{L} = (\mathbf{l}_1, \dots, \mathbf{l}_n)$ defines the configuration of each part and \mathbf{C}_i refers to the pixels covered by part i . $p(\mathbf{L})$ is defined as in Section 2.1 and places plausible limits on the joint angles of the hands relative to the lower arms.

The appearance term, $p(\mathbf{C}_i|\mathbf{l}_i)$, is composed of the product of pixel likelihoods using colour distributions modelled by mixtures of Gaussians, and edge and illumination cues added through HOG descriptors.

Sampling from Eqn. 2 is facilitated by the restriction to tree-like topologies and can as a result be performed in time linear in the number and configurations of parts (Felzenszwalb and Huttenlocher, 2005).

Improvements in sampling efficiency. When using a sampling method to propose plausible arm locations, it is important that the true arm configuration is contained in the set of samples. In this respect the tree-structured pictorial structure sampler is insufficient; for example, given an image where a part is partially or completely occluded, the associated probability for this part to be generated from its true location can be very low. To increase the probability of

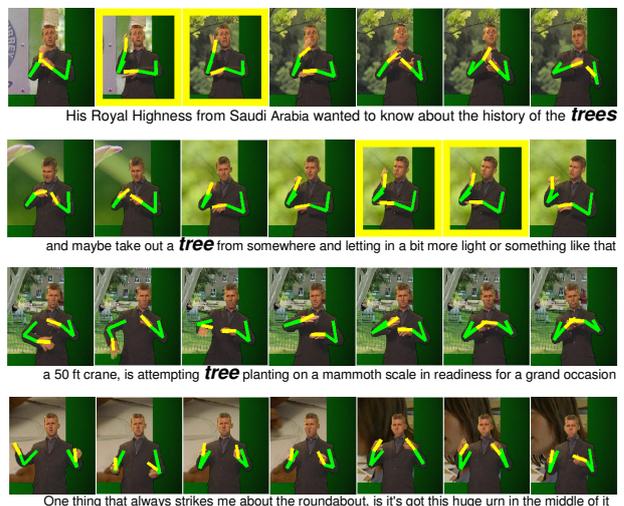


Figure 4: **Example training data for the target sign ‘tree’.** The top three rows are positive subtitle frame sequences (each around 20 seconds long), selected because they contain the text word ‘tree’. However, the sign only appears in the first two (outlined in yellow). The final row is an example negative subtitle sequence which does not contain the text word ‘tree’ and also does not, in fact, contain the sign for tree. Signs are learnt from such weakly aligned and noisy data.

sampling the true configuration, we propose several modifications in Buehler et al. (2008), such as sampling from the max-marginal instead of the marginal distribution which is typically used.

3 Automatic sign learning

This section outlines our approach for automatically learning signs from signed TV broadcasts. We describe how *weak* supervision is extracted from subtitles, visual description and matching of signs, and a multiple instance learning method for learning a sign despite the weak and noisy supervision. A more detailed discussion of the method can be found in Buehler et al. (2009).

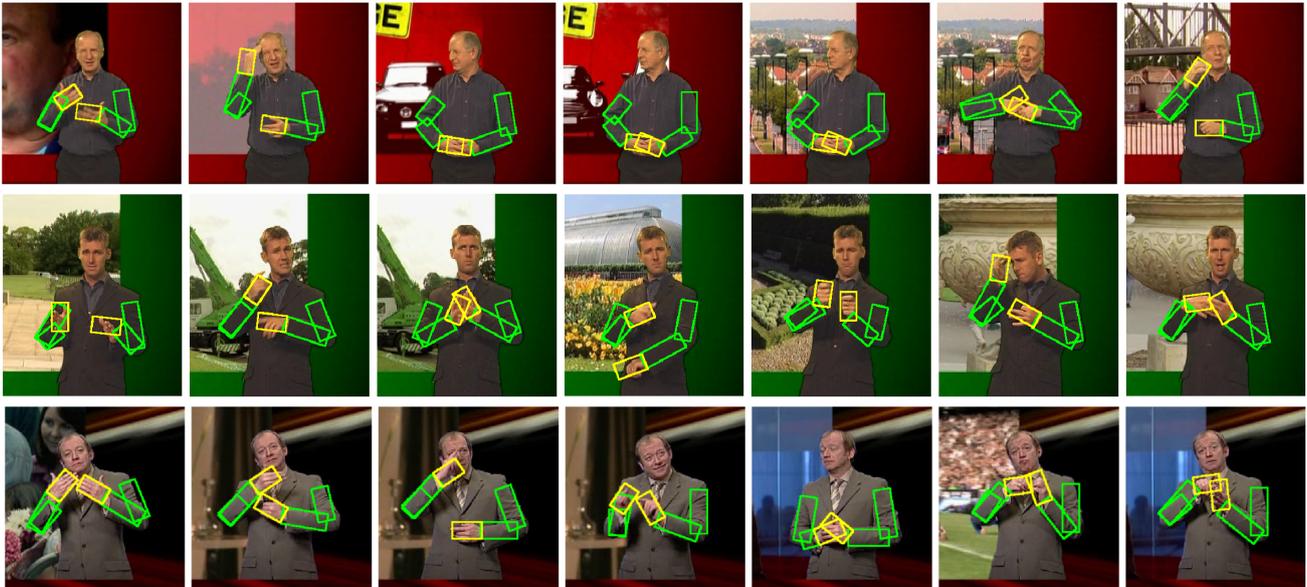


Figure 5: **Sample of tracking results on hour-long sequences.** The estimated pose is shown for uniformly spaced frames in three hour-long sequences with different signers. The pose is qualitatively highly accurate in all frames.

3.1 Automatic generation of training data

By processing subtitles we can obtain a set of video sequences labelled with respect to a given target English word as ‘positive’ (likely to contain the corresponding sign) or ‘negative’ (unlikely to contain the sign); this is illustrated in Figure 4. Hand detection using our articulated upper body tracker (Section 2), and feature extraction are then applied to extract visual descriptions for the sequences.

To reduce the problems of polysemy and visual variability for any given target word we generate training data from the same signer and from within the same topic (*e.g.* by using a single TV program). Even when working with the same signer, the intra-class variability of a given sign is typically high due to ‘co-articulation’ where the preceding or following signs affect the way the sign is performed, expression of degree (*e.g.* ‘very’) or different emotions, and varying locations relative to the body.

3.1.1 Text processing

Subtitle text is extracted from the recorded digital TV broadcasts by simple OCR methods (Everingham et al., 2006) (UK TV transmits subtitles as bitmaps rather than text). Each subtitle instance consists of a short text, and a start and end frame indicating when the subtitle is displayed. Typically a subtitle is displayed for around 100–150 frames.

Given a target *word* specified by the user, *e.g.* “golf”, the subtitles are searched for the word and the video is divided into ‘positive’ and ‘negative’ sequences.

Positive sequences. A positive sequence is extracted for each occurrence of the target word in the subtitles. The alignment between subtitles and signing is generally quite imprecise because of latency of the signer (who is translating from the soundtrack) and differences in word/sign order, so some ‘slack’ is introduced in the sequence extraction. Consequently, positive sequences are, on average, around 400 frames in length. In contrast, a sign is typically

around 7–13 frames long. This represents a significant correspondence problem.

The presence of the target *word* is not an infallible indicator that the corresponding *sign* is present – examples include polysemous words or relative pronouns *e.g.* signing “it” instead of “golf” when the latter has been previously signed. We measured empirically that in a set of 41 ground truth labelled signs only 67% (10 out of 15 on average) of the positive sequences actually contain the sign for the target word.

Negative sequences. Negative sequences are determined in a corresponding manner to positive sequences, by searching for subtitles where the target word *does not* appear. For any target word an hour of video yields around 80,000 negative frames which are collected into a single negative set. The absence of the target *word* does not always imply that the corresponding *sign* is not present in the negative sequences. This is because different words might be signed similarly, or a sign might be present in the video but not appear in the subtitles (*e.g.* referred to as “it”).

3.1.2 Visual processing

A description of the signer’s actions for each frame in the video is extracted by tracking the hands via our upper body model (Section 2). Descriptors for the hand position and shape are collected over successive frames to form a *window* descriptor which forms the unit of classification for learning. The temporal length of the window is between 7 and 13 frames, and is learnt for each sign.

Hand shape description. The ‘shape’ of the hands is extracted by segmentation, and represented by a HOG descriptor (Dalal and Triggs, 2005; Kjellström et al., 2008). HOG descriptors are chosen for their ability to capture both boundary edges (hand silhouette) and internal texture (configuration of the fingers), and the contrast normalization they employ gives some invariance to lighting.

To deal with cases where the hands are overlapping or touching, descriptors for each hand and also for the pair of hands are extracted in parallel.

3.2 Measuring visual distance between signs

Our learning approach seeks temporal *windows* of video which represent the same sign, where a window is the concatenation of visual descriptors for a sequence of frames. In order to compare two such windows a distance function is needed which captures differences in position and motion of the hands and their appearance.

For each frame t of the window, each hand is described by a vector $\mathbf{x}(t) = \langle \mathbf{x}_{pos}, \mathbf{x}_{dez}, \mathbf{x}_{dezP} \rangle$ which combines hand position (pos) and shape (dez) for both the individual hand and the combined hand pair (subscript P). The descriptor for a window \mathbf{X} is the concatenation of the per-frame descriptors $\mathbf{x}(t)$.

In BSL one hand is dominant, while the position and appearance of the other hand is unimportant for some signs. We build this into our distance function. Given two windows \mathbf{X} and \mathbf{X}' the distance between them is defined as the weighted sum of distances for the right (dominant) and left (non-dominant) hands:

$$D(\mathbf{X}, \mathbf{X}') = d_R(\mathbf{X}, \mathbf{X}') + w_L d_L(\mathbf{X}, \mathbf{X}') \quad (3)$$

where $d_L(\cdot)$ and $d_R(\cdot)$ select the descriptor components for the left and right hands respectively. The weight $w_L \leq 1$ enables down-weighting of the non-dominant hand for signs where it does not convey meaning. We refer to two windows \mathbf{X} and \mathbf{X}' as showing the same sign if their distance $D(\mathbf{X}, \mathbf{X}')$ is below a threshold τ . Section 3.3 describes how w_L and τ is learnt for each individual target sign.

The distance measure for the left and right hand alike is defined as a weighted sum over the distances of the position, shape and orientation components (we drop the hand subscript to simplify notation):

$$d(\mathbf{X}, \mathbf{X}') = w_{pos} d_{pos}(\mathbf{X}, \mathbf{X}') + w_{dez} d_{dez}(\mathbf{X}, \mathbf{X}') + w_{ori} d_{ori}(\mathbf{X}, \mathbf{X}') \quad (4)$$

The hand shape distance d_{dez} is computed with invariance to rotation. This is in accordance with linguistic sign research (Brien, 1993), where different hand configurations are described separately by shape (d_{dez}) and orientation (d_{ori}). The position distance d_{pos} is designed to be invariant to small differences in position, since repetitions of the same sign can be performed at different positions (*e.g.* this applies especially to signs performed in front of the chest). For a detailed description of these distance functions see Buehler et al. (2009).

The positive weights w_{pos} , w_{dez} and w_{ori} are learnt off-line from a small number of training examples.

3.3 Automatic sign extraction

Given a target word, our aim is to identify the corresponding sign. The key idea is to search the positive sequences to provide an example of the sign. Each positive sequence in turn is used as a ‘driving sequence’ where each temporal window of length n within the sequence is considered

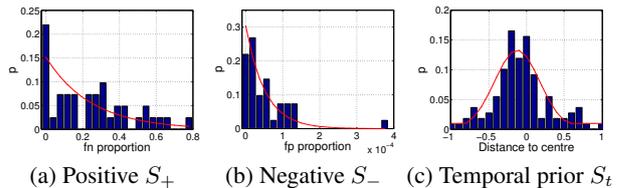


Figure 6: **Distributions used to score template windows.** Plots (a) and (b) show the empirical distribution of errors (bars) and the fitted exponential distribution (curve). Note the scale on the x -axis. Plot (c) shows the temporal distribution of signs within corresponding positive sequences.

as a *template* for the sign. We require a score function to evaluate each template, with a high score if the template occurs within most of the positive sequences and not often in the negative data. The sign is determined by maximizing the score function over all templates, over the sign specific dominant/non-dominant hand weighting w_L , and over the threshold τ which indicates if two signing windows show the same sign..

Multiple instance learning method. For a hypothesized setting of the classifier parameters $\theta = \{\hat{\mathbf{X}}, w_L, \tau\}$, with template window $\hat{\mathbf{X}}$ of length n , we assign a score

$$S(\theta) = S_+(\theta) + S_-(\theta) + S_t(\theta) \quad (5)$$

to the classifier as a function of (i) its predictions on the positive sequences S_+ and the negative set S_- , and (ii) our prior knowledge about the likely temporal location of target signs S_t .

Unfortunately, when designing S_+ and S_- , we know that some non-negligible proportion of our ‘ground truth’ labels obtained via the subtitles will be incorrect, *e.g.* in a positive sequence the target word appears but the corresponding sign is not present, or in the negative data the target sign is present but not the corresponding target word. A model of such errors is empirically learned and approximated using exponential models (see Figure 6a,b).

The sign instances which correspond to a target word are more likely to be temporally located close to the centre of positive sequences than at the beginning or end. As shown in Figure 6c, a Gaussian model gives a good fit to the empirical distribution. The temporal prior p_t is learnt from a few training signs as for the score functions.

Searching for the sign by maximizing the score. Given a template window $\hat{\mathbf{X}}$ of length n from a positive sequence, the score function is maximized using a grid search over the weight for the left hand w_L , and over a set of similarity thresholds τ . This operation is repeated for all such template windows, with different lengths n , and the template window that maximizes the score is deemed to be the sign corresponding to the target word.

Using a per-sign window length allows for some signs being significantly longer than others. The weight w_L allows the importance of the left hand to be down-weighted for signs which are performed by the right hand alone.

3.4 Experiments

Given an English word our goal is to identify the corresponding sign. We deem the output a success if (i) the selected template window, *i.e.* the window with the highest score, shows the true sign (defined as a temporal overlap of at least 50% with ground truth) *and* (ii) at least 50% of all windows within the positive sequences which match the template window show the true sign.

Datasets. We tested our approach on 10.5 hours of signing sequences recorded from BBC broadcasts (including subtitles), aired between 2005 and 2006, and covering such diverse topics as politics, motoring and gardening. Signing is performed by three different persons. The image size after cropping the area around the signer is 300×330 pixels.

Test set. The method is evaluated on 210 words. These words were selected and fixed before running the experiments, without knowledge of the appearance of the target signs, *i.e.* how the corresponding sign is performed. Selection was based on: (i) the target word must occur more than 5 times in the subtitles; (ii) the target word is a verb, noun or adjective as opposed to linking words such as “then”, “from”, “the”, etc.; (iii) the target word does not have multiple meanings (as opposed to *e.g.* the word “bank”).

The full list of signs used is given at www.robots.ox.ac.uk/~vgg/research/sign_language/, which also contains example sequences of the detected signs.

Results. In 136 out of 210 cases (65%) we are able to automatically find the template window which corresponds to the target sign (see Figure 1 for two examples).

The precision-recall curve in Figure 7 (blue dashed line) shows that the score associated with a template window can be used as a confidence measure, giving an extremely good guide to success: at 11% recall (23 signs) precision is 100%; at a recall of 50% (105 signs) the precision is 77%.

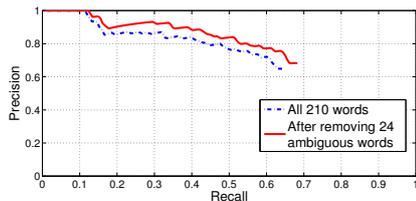


Figure 7: **Precision recall curve** computed using the score of the template window to rank learned signs.

Some words in our dataset co-occur with other words in the subtitles *e.g.* “prince” and “charles”, which renders the correct template window ambiguous. Often these incorrectly-learned signs have a high associated score and hence reduce the precision even at low recall. By using simple statistics we can exclude 24 words from processing which leads to an improved precision-recall curve in Figure 7 (red solid line). We achieve good results for a variety of signs: (i) signs where the movement of the hand is important *e.g.* “golf”, (ii) signs where the hands do not move but the hand shape is important *e.g.* “animal”; (iii) signs where both hands are together to form a sign *e.g.* “plant”; (iv) signs which are finger spelled *e.g.* “bamboo”; (v) signs which are performed in front of the face *e.g.* “visitor”, which makes identifying the hand shape difficult.

Some of the mistakes are understandable: For the word “wood”, our result is the sign for “fire”. This is not sur-



Figure 8: **Challenges for signer-independent sign recognition.** Repetitions of the same sign can differ in the hand position, the hand movement, the hand shape, and even in the number of hands involved. This is illustrated for three instances of the sign “bad”.

prising since these two words often appeared together. The sign “year” is difficult since the signs for “last year”, “this year” and “next year” differ – our method picks the sign for “next year”.

4 Signer-independent sign recognition

Having demonstrated a method for learning a sign automatically, it is natural to investigate if the learnt sign can be recognized across different signers. The problem is that the learning method of Section 3 is built on the restrictions that apply to a single signer, for example that the lighting, body size and position do not vary significantly and also that (apart from co-articulations) the same sign is performed in a consistent style. These restrictions do not apply when the signer changes (see Figure 8) – signs can be performed with different speeds, variable extents, at varying locations, or with slightly different hand shapes. In many cases, these variations are due to the differences in signing between different people, such as local accents, or personal traits. The visual features and restrictions of Section 3 which took advantage of the single-signer situation are not sufficient for signer-independent recognition.

However, in this section we demonstrate that by adding a discriminative training stage signs *can* be recognized and localized in new signers. The experiments illustrate the extent to which our features (hand trajectory and hand shape) generalise to previously unseen signers.

4.1 Method

Our goal here is to *detect* signs in previously unseen signers using the automatically learnt signs from Section 3 within a temporally restricted search space. That is, instead of detecting a given sign in a full TV show (1 hour long), we search for it within short “positive sequences” extracted from around the word occurrences of the corresponding English word in the subtitles. In Section 3.3, a temporal prior is used to favour sign occurrences near the centre of a positive sequence. In this section, instead of such a prior, we use smaller positive sequences (on average half the length; 10 seconds long) extracted with a small offset from the word occurrence to take the empirically observed latency

between subtitles and signing into account (see Figure 6c). Note that empirically the sign we aim to detect is only performed in around half of the positive sequences, since the occurrence of a subtitle word does not imply the presence of the corresponding sign (see Section 3.1.1; although here, the positive sequences are shorter). Therefore, even a perfect sign detector would have an accuracy of at most 50%. Assume that we have learnt a sign from a *training signer* using the method of Section 3. For a learnt sign we have an automatically learned template window \hat{X} with highest score, and all windows which are similar to the template, *i.e.* with a distance to \hat{X} below a threshold τ of the pairwise distance measure. We consider these windows as a positive training set.

We compare two methods for generalizing from the sign learnt from the training signer to recognizing this sign for other signers. (i) **Template matching:** The pairwise distance measure from Section 3.2 is used as classifier to identify signs which are similar to the learnt template window \hat{X} . (ii) **Discriminative training:** A support vector machine (SVM) classifier is trained to detect the sign. Training data consists of the positive examples from the training signer, and negative examples taken from all the other signs considered.

For a given English word, a positive sequence for each word occurrence in the subtitles is extracted. Our aim is to find the corresponding sign in each of these sequences. This is achieved in a sliding window fashion by searching for the window with highest confidence according to (i) the corresponding SVM output or (ii) similarity to the learnt template. We search over different window lengths, since the duration of a sign in a positive sequence is unknown. In this way, one window is selected from each sequence.

Features. We use information from two cues: hand position and hand shape for each frame as described in Section 3. The hand shape cue is based on a set of hand exemplars which are used to describe the hand shape for each frame (think visual words); see Buehler et al. (2009) for a detailed description. The position of the left eye is automatically detected in each frame using the method of Everingham et al. (2006) and serves as reference point. Each training sample is down-sampled to be of equal temporal length (5 frames) – a prerequisite for SVM training.

SVM classifier. We use the LIBSVM library (Chang and Lin, 2001) to learn a binary SVM for each of the 15 different words in our dataset (see Section 4.2). Separate Radial Basis Function (RBF) kernels are computed for the hand position and the hand shape cue individually, and combined by computing the mean. We also evaluated using the product over the individual kernels instead, which gave comparable performance.

4.2 Experiments

Dataset. Experiments are performed for 15 English words, selected such that each word occurs more than five times in the subtitles of a specific signer (the training signer). The selected words are: better, Britain, car, help, hope, house, kitchen, money, mourn, new, night, room, start, team, and week.

For these 15 words, signing examples from the training signer are automatically extracted using our method from Section 3, and used to train initial SVM classifiers. Note that this includes wrongly learned signs, as is shown in Table 1, column “Learning - FP”. These initial classifiers are subsequently used to extract additional signing examples from a database of 1730 positive sequences from 6 previously unseen signers (none of them being the training signer), each with a duration of 10 seconds. In this way, between 58 and 192 sequences are extracted for each word (Table 1, column “WS”).

Results. First, we automatically learn for each of the 15 English words the corresponding sign. Even though the supervision provided by the subtitles is very weak and noisy, our results are highly accurate: out of 195 automatically extracted signing examples, 164 are correct (see Table 1, column “Learning”). Note that only the sign for “team” is not learned correctly (0 true positives TP, but 5 false positives FP).

From this dataset, an SVM classifier is learned for each word, and subsequently used to detect one example of the corresponding sign within each of the positive sequences. We define a ranking of the detections by confidence based on (i) the SVM decision value of the detected sign, and (ii) the margin between (i) and the second highest decision value within the same sequence (using non-maximum suppression). We know that the sign is only performed in about half the positive sequences (Section 4.1), hence assuming that our detector finds a sign a little less than half of the time (if it is performed), then the 20% highest ranked detections should often be correct. Indeed, for this subset, on average 67% of the detections show the true sign (see Table 1, column “SVM detector”).

We further analysed the performance of our ranking function by plotting the proportion of correctly detected signs as a function of the highest ranked detections (Figure 9, blue curve).

The SVM classifiers used above were trained from automatically extracted signing examples, including 31 examples which do not show the correct sign (see Table 1, column “Learning - FP”). We observe a slight increase in accuracy if these examples are excluded from training (Figure 9, green solid curve).

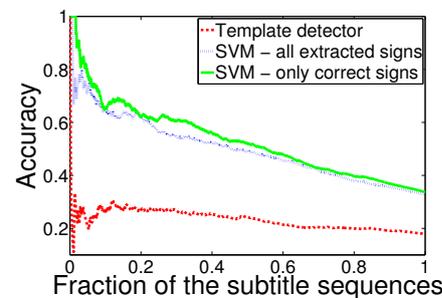


Figure 9: **Sign detection accuracy.** For each classifier the accuracy is shown as a function of detection confidence.

Comparison to template detector. We repeat the sign detection results from this section, using as a classifier the template-based distance function instead of the SVM approach. The results clearly deteriorate, as can be seen in Figure 9 (red dotted curve) and in Table 1 (column “Template detector”).

Sign	WS	Learning			SVM detector				Template detector		
		TP	FP	Ratio	TP	FP	Ratio	SGR	TP	FP	Ratio
better	95	11	2	0.85	2	2	0.50	2	2	15	0.12
Britain	98	9	1	0.90	18	13	0.58	3	1	17	0.06
car	63	15	1	0.94	6	4	0.60	4	15	4	0.79
help	150	22	1	0.96	6	3	0.67	3	15	20	0.43
hope	61	7	1	0.88	8	7	0.53	4	7	13	0.35
house	177	7	5	0.58	30	15	0.67	3	21	43	0.33
kitchen	58	7	0	1	6	0	1	3	2	24	0.08
money	102	6	1	0.86	10	3	0.77	3	7	15	0.32
mourn	77	8	0	1	10	2	0.83	2	4	11	0.27
new	183	20	4	0.83	47	7	0.87	4	2	13	0.13
night	62	8	0	1	11	0	1	3	3	4	0.43
room	126	9	0	1	10	0	1	2	2	2	0.50
start	192	17	10	0.63	26	67	0.28	5	5	5	0.50
team	151	0	5	0	0	4	0	2	0	41	0
week	135	18	0	1	20	6	0.77	2	6	23	0.21
MEAN	115			0.83			0.67	3.0			0.30

Table 1: **Recognizing signs in new signers.** For 15 English words, the corresponding signs are automatically learned (Section 3), and then used to recognize the sign in new signers (Section 4). Column “WS” shows the number of positive sequences for each word. For our automatic sign learning method, the number of correctly learned signs (TP), incorrectly learned signs (FP), and the ratio $TP/(FP+TP)$ is given (column “Learning”). Subsequently, additional signing examples are detected within the 1730 positive sequences, either using the SVM framework as described in this section (column “SVM detector”), or the automatically found sign templates for each word (see Section 3) as detectors (column “template detector”). The number of new signers for which signing examples are extracted is given in column “SGR”. Note that the values in the columns “SVM detector” and “Template detector” are computed using the 20% highest ranked sign detections (see also Figure 9, blue and red curves).

5 Conclusion

We described methods for visual tracking of a signer in complex TV footage, and for automatic learning of signs. The framework enables learning a large number of BSL signs from TV broadcasts using only supervision from the subtitles. We achieve very promising results even under these weak and noisy conditions.

We illustrated that examples automatically extracted for a single signer can be used to recognize a given sign for other signers provided an additional discriminative training stage is applied. This demonstrates that our features (hand trajectory and hand shape) generalise well across different signers, despite the significant inter-personal differences in signing.

Future work will concentrate on improving the accuracy of signer-independent recognition to a complete unconstrained scenario where no subtitles are available.

Acknowledgements. We are grateful for financial support from EPSRC, the Royal Academy of Engineering, ERC VisRec, and ONR MURI N00014-07-1-0182.

6 References

- D. Brien. 1993. *Dictionary of British Sign Language*.
- P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. 2008. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*.
- P. Buehler, M. Everingham, and A. Zisserman. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. CVPR*.
- C. C. Chang and C. J. Lin. 2001. LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- H. Cooper and R. Bowden. 2007. Large lexicon detection of sign language. *IEEE Workshop on Human Computer Interaction*.
- H. Cooper and R. Bowden. 2009. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *Proc. CVPR*.
- N. Dalal and B. Triggs. 2005. Histogram of oriented gradients for human detection. In *Proc. CVPR*.
- M. Everingham, J. Sivic, and A. Zisserman. 2006. Hello! My name is... Buffy – automatic naming of characters in TV video. In *Proc. BMVC*.
- A. Farhadi and D. Forsyth. 2006. Aligning ASL for statistical translation using a discriminative word model. In *Proc. CVPR*.
- A. Farhadi, D. Forsyth, and R. White. 2007. Transfer learning in sign language. In *Proc. CVPR*.
- P. Felzenszwalb and D. Huttenlocher. 2005. Pictorial structures for object recognition. *IJCV*, 61(1).
- E. J. Holden, G. Lee, and R. Owens. 2005. Automatic recognition of colloquial Australian sign language. In *Workshop on Motion and Video Computing*.
- T. Kadir, R. Bowden, E. J. Ong, and A. Zisserman. 2004. Minimal training, large lexicon, unconstrained sign language recognition. In *Proc. BMVC*.
- H. Kjellström, J. Romero, D. Martínez, and D. Kragić. 2008. Simultaneous visual recognition of manipulation actions and manipulated objects. In *Proc. ECCV*.
- S. Nayak, S. Sarkar, and B. Loeding. 2009. Automated extraction of signs from continuous sign language sentences using iterated conditional modes. In *Proc. CVPR*.
- T. Starner, J. Weaver, and A. Pentland. 1998. Real-time American sign language recognition using desk- and wearable computer-based video. *IEEE PAMI*, 20(12).

Towards Czech on-line sign language dictionary – technological overview and data collection

Pavel Campr¹, Marek Hruz¹, Jiří Langer², Jakub Kanis¹, Miloš Železný¹, Luděk Müller¹

¹ Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic

² Institute of Special Education Studies, Faculty of Education, Palacký University in Olomouc, Czech Republic
campr@kky.zcu.cz, mhruz@kky.zcu.cz, jiri.langer@upol.cz,
jkanis@kky.zcu.cz, zelezny@kky.zcu.cz, muller@kky.zcu.cz

Abstract

In this article we present the current state of our work on an on-line sign language dictionary. The aim is to create both an explanatory and a translation dictionary. It is primarily targeted (but not limited) to the Czech and Czech sign language. At first we describe technological aspects of the dictionary and then our data collection practices. The dictionary is an on-line application build with respect to the linguistic needs. We use written text to represent spoken languages and several representations are supported for sign languages: videos, images, HamNoSys, SignWriting and interactive 3D avatar. To decrease time required for data collection and publishing in the dictionary we use computer vision methods for video analysis to detect sign boundaries and analyze the manual component of performed sign for automatic categorization. The content will be created by linguists using both new and already existing data. Then, the dictionary will be opened to the public with possibility to add, modify and comment data. We expect that this possibility of on-line elicitation will increase the number of informants, cover more regions and makes the elicitation cheaper and the evaluation easier. Furthermore we prepare a mobile interface of the dictionary. The mobile interface will use different format of web pages and different video compression methods optimized for slower Internet connection. We also prepare an offline version of the dictionary which can be automatically generated from the online content and downloaded for offline usage.

1. Introduction

As for spoken languages, sign languages can utilize dictionaries for several purposes. Translation dictionaries are used to translate words (or phrases) from one language to another, explanatory dictionaries define the words in the same language instead of translating them. Traditional dictionaries for spoken languages use written text as main form for content creation. This becomes more difficult for sign languages where written form of the language is not so evolved and spread among the community. Examples of the written forms are HamNoSys (Hamburg Sign Language Notation System, developed in 1985) and SignWriting (developed in 1974). Advances in the field of information technologies allow creation of electronic dictionaries with new possibilities such as interactivity, faster searching, video animations, etc. The Internet brought new platform for on-line applications which opened other possibilities for the dictionaries: availability from anywhere, information sharing, interoperability and replaced static content with dynamic.

There are many existing on-line sign language dictionaries, but not all of them offer expected features, quality and content:

- Easy and intuitive usage
- Searching - not only by text, but by another, sign language specific criteria
- Complete data - to cover whole language, not only limited topics
- Being up-to-date

- Usage of the sign language written forms (HamNoSys, SignWriting, etc.)
- Linguistic information
- Version for mobile devices
- Offline version for download

Usually the existing dictionaries are specialized for selected topics or support only limited features. Our goal is to create state-of-the-art sign language dictionary which supports all mentioned features, is both translational and explanatory, and supports unlimited number of languages so that a dictionary entry (word or collocation) can be translated e.g. from the Czech sign language to the Czech language, English and American Sign language. Our dictionary is being developed now and our main content will be the Czech sign and Czech language.

spoken languages	sign languages
written text	HamNoSys SignWriting video image or illustration 3D avatar

Table 1: Content forms supported by the dictionary

The dictionary supports several forms of content for sign languages as seen in the table 1. Along the common forms an interactive 3D avatar is available. It can perform the sign and the user can change angle of the view, zoom to a

detailed parts and slow down or pause the animation. Another innovation is usage of computer vision methods for video analysis to speed up video data collection and automatically categorize the signs into groups, which can be used as a criterion for searching.

This article is divided into four main parts: *Dictionary structure and content* describes, which data can be stored and viewed in the dictionary. In *Requirements for dictionary usage*, three different platforms (PC with and without the Internet and mobile devices) and their requirements are presented. *Searching* part describes possibilities of searching in the dictionary database. Finally, *Data collection* part introduces our practises for data collection and usage of computer vision methods.

2. Dictionary structure and content

Our on-line dictionary supports unlimited number of languages, the content can be represented by all forms listed in the table 1 and the dictionary entries can contain explanatory part and translations to other languages.

The functionality is based upon a database structure which is shown on the figure 1.

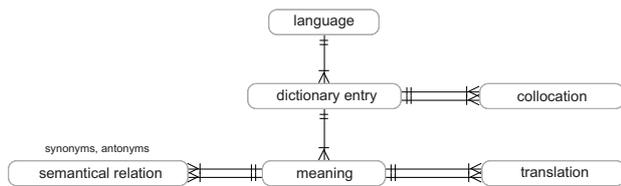


Figure 1: Database structure (simplified view)

Every dictionary entry is connected to one language. The entry can have one or more meanings. An explanation, use cases and linguistic information (e.g. part of speech) are attached to the meaning.

As collocations are widely used in the Czech sign language the database allows linkage between a word from a collocation to the dictionary entry, which corresponds to the stand-alone word. For example, collocation *personal baggage* is linked to *personal* and *baggage* entries.

Each of the meanings can be attached to another one (of a different dictionary entry) and express synonymic or antonymic relation.

The translation functionality is allowed by the linkage between two meanings which are linked to entries in different languages. This means that each of the meanings can be translated separately.

This database structure fulfils all linguistic needs for creation of explanatory and translation dictionary.

Every form specified in the table 1 is equal to another one and are optional (at least one of them is mandatory). Here we discuss some specific features:

video The dictionary entry can be represented by one or several video clips. They can be recorded separately (different speaker, place) or can be recorded simultaneously from multiple views (e.g. front, side and face view). The video data can be stored on the dictionary server or can be

stored on external server anywhere on the Internet, including video sharing websites (Youtube, Vimeo, Dailymotion etc.). The video data will be available in more compression qualities and sizes, mainly for the usage on mobile devices.

image Multiple illustrations, photos or any other images with representation of a sign can be used.

HamNoSys (Hamburg Sign Language Notation System)

The dictionary includes a special editor for HamNoSys strings which allows the users to create new or modify existing HamNoSys strings.

SignWriting Similarly to HamNoSys, special editor for SignWriting is required. This editor is being developed now and will be added lastly.

3D avatar Synthesis of the sign language creates a computer animation of the signing avatar (see fig. 2). For this purpose, we have specially created 3D animation model of the human figure.

For the web environment we had to convert the animation model to Collada format¹. This format allows us to save 3D data, define the skeleton avatar animation and import the control trajectories.

An important part of the synthesis system is a conversion algorithm which converts a symbolic entry into the control instructions that are transmitted to the animation model. The entry of the algorithm is one or more signs noted in HamNoSys. The conversion algorithm was originally designed for the manual component of the sign language (Křňoul et al., 2008) and the version HamNoSys 3. We can convert not only isolated signs but the phrase, or continuous speech. The initial perceptual study shows good clarity of the animation of the manual component. The non-manual component was initially expressed by the visual speech, i.e. the articulation of words spoken language (mouthing). New extension of the conversion algorithm, however, allows transfer of the non-manual signals (NMS). For this purpose, a methodology for notation of NMS is designed (Křňoul, 2010). The notation of NMS is now a part of signs and the user is allowed to edit movements of torso, head and facial expressions.

An animation from the symbols has the benefits from the possibility of easy editing signs. The user can change the notation and determine the best form of the sign. One sign may be used for creation another sign with a similar form. The synthesis system provides two types of interactivity for the dictionary purpose. The first type is a preview of the figure. The animation model is rendered in the window and user can turn it in three axes or zoom the facial details, etc. Unlike video, which is always defined in one direction, the user can adjust it for best view. The animation is not in principle blurred or noisy. The second type of interactivity is phasing of the animation. User can suspend animation, re-run or step frame by frame. In particular, stepping allows the user to find "an articulatory target". The articulatory target is shape and position of the hand, body posture or facial expression that establishes meaning of the sign. From the educational point of view it allows the users quickly understand and learn new sign.

¹OpenCOLLADA Framework, www.collada.org



Figure 2: 3D avatar. Left: front view. Right: face expression.

3. Requirements for dictionary usage

Primary platform for the dictionary usage is a PC connected to the Internet, with any modern internet browser (Internet Explorer version 6 and above, Firefox, Opera, Chrome, etc.) with installed Adobe Flash plugin. To enable hardware accelerated, high quality 3D avatar animation, a special plugin (Google O3D) is required. Without this a lower quality 3D avatar is used, without hardware acceleration. In the future, WebGL (new specification for writing web applications utilizing hardware accelerated 3D graphics) can be used.

Secondary platform is a mobile device (PDA, smartphone, etc.) with installed internet browser. For this platform the dictionary will be formatted with respect to the device capabilities and the video clips will be resized and compressed for the needs of those devices.

Another secondary platform is a PC in the same configuration as above but without internet connection. The dictionary will be able to automatically create offline version, which will be automatically created every day and available for download. This offline version will be limited in functionality in comparison to online version, mainly in searching capabilities.

4. Searching

Key feature of the dictionary is searching. The goal is to create searching functionality which will provide relevant results for user query. For spoken languages the user provides searched term, language and optionally a topic and grammatical information (e.g. part of speech). Result is a list of dictionary entries which satisfy the given search criteria. For the Czech language a lemmatization engine is used to enable searching among different inflected forms of same words. Furthermore, the searching is not limited only for dictionary entry title but provides fulltext search in all text items (meanings, explanations, use cases etc.).

For sign languages the searching feature is more complicated since the sign language words and sentences aren't represented in text form and thus we cannot use tools used for text searching. Our goal is to examine possibility of HamNoSys and SignWriting usage as search criteria and find a way how to find related dictionary entries for the

given criteria. Because we expect that the resulting list for this way of searching will contain many items, other criteria can be used to limit the search as for text search (topic, grammatical information).

5. Data collection

The content of the dictionary will be continuously extended and modified. For this purpose a special administration section is available where the users can (depending on their permissions) create, update or delete dictionary entries. Special workflow management is prepared for administration users with limited permissions, where all modifications must be confirmed by administrators with full permissions. Thus the quality of the content is preserved with the possibility for many users to edit the content. The workflow can be easily changed after we get some experience after the dictionary is released.

The decision if a new or updated dictionary entry is valid will be supported by a discussion under each dictionary entry, where the community can decide, whether the provided information is correct.

Most of the dictionary entries will be provided by professional linguists. The process of video recording is quite time consuming, to reduce the required time we use several tools, such as automatic detection of sign boundaries in a recorded session.

5.1. Utilization of computer vision methods

We use computer vision techniques to automatically detect boundaries of signs in a recorded session. There are certain conditions that need to be met in order to successfully obtain the boundaries. There should be a neutral pose of the signing person. This pose defines the beginning and the end of the sign. Also, the stage where the person is signing should have laboratory-like conditions so that the hands of the person are clearly visible and easily distinguishable from the background. Since the intention of the recordings is to use them in a SL dictionary these conditions are rational.

5.2. Sign boundaries detection



Figure 3: Video file processing and segmentation. Left: original frame from a video. Right: segmented image, white parts correspond to head and hands.

We detect two features: motion and position. First, the image is pre-processed and segmented so that we obtain parts of human body. In some cases a simple thresholding can be used (e.g. the signer wears dark clothes), see fig. 3. In

more complex situations when the brightness level of pixels is not enough to distinguish between parts of human body and the rest of the scene, we use skin color segmentation. Next, we use object detection in the segmented image. We compare the position of objects (hands and head) with the trained initial position. If the distance is below a threshold we assume the signer is in the initial pose (fig. 4). In some cases we do not need to compute the distances but rather examine the position of the object and check whether it is in some predefined region. This is just an alternative approach with the same mathematical foundation.

In the next step we describe the movement as the sum of pixels in the difference image. This does not give us a detailed description of the movement but rather an estimate of total movement in the image. This value is normalized by the resolution of the image. A threshold is set and when the relative motion in the image is above this threshold we assume there is a significant movement of hands (fig. 4).

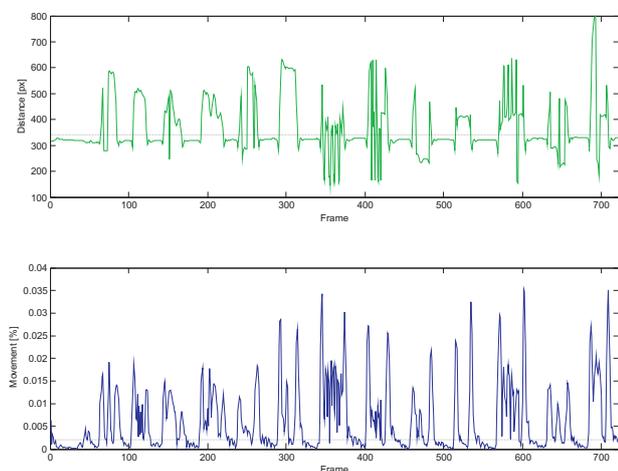


Figure 4: Features used for sign boundaries detection: motion and position

The features of movement and initial pose are measured over the recording. The first frame in which the neutral pose is not detected and movement is detected is considered as the beginning of the sign. Respectively, the first frame in which the neutral pose is detected and no movement is detected is considered to be the end of a sign. We have to shift the boundaries of the detected sign a bit so that the resulting cuts begin and end in a stationary pose. Usually we use the value of ± 50 ms.

5.3. Automatic processing of signs

According to work described in Hrůz et al. (2008) we are able to track hands and head in recordings designed for sign language dictionary. For now we are able to obtain the trajectories of both hands and the head. On a relatively small dataset (Campr et al., 2007) we achieved good recognition results (Trmal et al., 2008) with features describing the manual component of SL. It is a baseline system and the features can be used for the annotation of a portion of manual component in the desired form. One of our goals is to develop a new system capable of describing the manual

component in more detail. Based on that we can automatically group similar signs and utilize this information for searching purposes.

6. Conclusion

In this article we presented our progress on ongoing project *Czech on-line sign language dictionary*. Some parts of the dictionary are nearly finished (database system, administration interface, 3D avatar, video players, fulltext search), other parts are being developed (sign search, SignWriting editor, frontend interface). We expect the first public release (and first feedback from the users) in the second half of the year 2010, but some limited pre-release versions are already available.

Our goals are both to create high-quality on-line sign language dictionary system and to provide high-quality content for the Czech and Czech sign language.

7. Acknowledgement

This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416, by the EU and the Ministry of Education of the Czech Republic, project No. CZ.1.07/2.2.00/07.0189, by the Ministry of Education of the Czech Republic, project No. ME08106 and by the Ministry of Education of the Czech Republic, project No. MŠMT LC536. We would like to thank Dr. Petr Peňáz, Masaryk University, Brno, for helpful discussions about linguistic aspects of the dictionary.

8. References

- Pavel Campr, Marek Hrůz, and Miloš Železný. 2007. Design and recording of signed czech language corpus for automatic sign language recognition. *Interspeech*, pages 678–681.
- Marek Hrůz, Pavel Campr, and Miloš Železný. 2008. Semi-automatic annotation of sign language corpora. *LREC Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pages 78–81.
- Zdeněk Krňoul, Jakub Kanis, Miloš Železný, and Luděk Müller. 2008. Czech text-to-sign speech synthesizer. *Machine Learning for Multimodal Interaction, Series Lecture Notes in Computer Science*, 4892:180–191.
- Zdeněk Krňoul. 2010. New features in synthesis of sign language addressing non-manual component. In *4th Workshop on Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. ELRA.
- Jan Trmal, Marek Hrůz, Jan Zelinka, Pavel Campr, and Luděk Müller. 2008. Feature space transforms for czech sign-language recognition. *Interspeech*, pages 2036–2039.

MobileASL: Overcoming the technical challenges of mobile video conversation in sign language

Anna C. Cavender¹, Neva Cherniavsky¹, Jaehong Chon², Richard E. Ladner¹,
Eve A. Riskin², Rahul Vanam², Jacob O. Wobbrock³

¹Computer Science & Engineering, ²Electrical Engineering, ³The Information School
University of Washington
Seattle, WA USA 98195

¹(cavender, nchernia, ladner)@cs.washington.edu

²(jaehong, riskin, rahulv)@ee.washington.edu

³wobbrock@u.washington.edu

Abstract

As part of the ongoing MobileASL project, we have built a system to compress, transmit, and decode sign language video in real-time on an off-the-shelf mobile phone. In this work, we review the challenges that arose in developing our system and the algorithms we implemented to address them. Separate parts of this research have been previously published in (Cavender et al., 2006; Cherniavsky et al., 2007; Cherniavsky et al., 2008; Vanam et al., 2009; Chon et al., 2009; Cherniavsky et al., 2009).

Compression and transmission of sign language video presents unique difficulties. We must overcome weak processing power, limited bandwidth capacity, and low battery life. We also must ensure that the system is usable; that is, that the video is intelligible and the algorithms that we employ to save system resources do not irritate users.

We describe the evolution of the MobileASL system and the algorithms we utilize to achieve real-time video communication on mobile phones. We first review our initial user studies to test feasibility and interest in video sign language on mobile phones. We then detail our three main challenges and solutions. To address weak processing power, we optimize the encoder to work on mobile phones, adapting a fast algorithm for distortion-complexity optimization to choose the best parameters. To overcome limited bandwidth capacity, we utilize a dynamic skin-based region of interest, which encodes the face and hands at a higher bit rate at the expense of the rest of the image. To save battery life, we automatically detect periods of signing and lower the frame rate when the user is not signing.

We implement our system on off-the-shelf mobile phones and validate it through a user study. Fluent ASL signers participate in unconstrained conversations over the phones in a laboratory setting. They find the conversations with the dynamic skin-based region of interest more intelligible. The variable frame rate affects conversations negatively, but does not affect the users perceived desire for the technology.

Ongoing work includes varying the spatial resolution instead of the temporal resolution, further optimization of rate-distortion-complexity, and a field study to determine usability over a long period of time in a realistic setting.

1. Introduction

Mobile technology has become an integral part of society, changing the nature of communication worldwide. The MobileASL project aims to expand accessibility for Deaf¹ people by efficiently compressing sign language video to enable mobile phone communication. Users capture and receive video on a typical mobile phone. They wear no special clothing or equipment, since this would make the technology less accessible.

There are three main challenges to building a system for real-time two-way video communication on mobile phones. First, the processing power on phones is weak. The encoder must run fast enough to show the video in real-time, and yet must produce intelligible video at low bit rates. Secondly, the bandwidth is limited. Video must be transmitted at rates of less than 30 kbps to be compatible with the capacity of the U.S. mobile phone network. Lastly, the battery capacity is low. Encoding, transmitting, receiving, and playing video

on a mobile phone quickly drains the battery, rendering the phone useless.

We develop sign language sensitive algorithms to attack these three challenges. We optimize the encoder parameters for the best possible tradeoff between efficiency and intelligibility, using an adaptation of a fast algorithm for distortion-complexity optimization. We address the problem of limited bandwidth by creating a dynamic skin-based *region-of-interest* (ROI) that encodes the face and hands at a higher bit rate at the expense of the rest of the image, increasing intelligibility without increasing bandwidth. We save power and processor cycles through automatic detection of periods of signing. When the user is not signing, we lower the frame rate, encoding and transmitting one tenth of the frames. We call this technique *variable frame rate* (VFR).

Our central goal is to increase access for Deaf people; we thus use intelligibility as our main measure of success. Throughout the evolution of our system, we verify our design and algorithms with users. We began the project by conducting focus groups and small laboratory studies to

¹Capitalized Deaf refers to members of the signing Deaf community, whereas deaf is a medical term.

validate our ideas. After building a working system, we evaluate it with a larger study in which fluent signers participate in unconstrained conversations over the phone.

1.1. Background

As is often the case with the design and implementation of a large system, separate parts of this research have been published previously (Cavender et al., 2006; Cherniavsky et al., 2007; Cherniavsky et al., 2008; Vanam et al., 2009; Chon et al., 2009; Cherniavsky et al., 2009). More complete versions of related work may be found in those publications.

Sign language video compression so that Deaf users can communicate over the telephone lines has been studied since at least the early 1980s. The first works attempted to enable communication by drastically modifying the video signal, e.g. by binarizing the image; (Foulds, 2006) provides a good overview. More closely related to our project are works that implement ROI encoding for reducing the bit rate of sign language video (Schumeyer et al., 1997; Woelders et al., 1997; Saxe and Foulds, 2002; Agrafiotis et al., 2003; Habili et al., 2004) and works that examine the intelligibility of sign language video at low frame rates (Sperling et al., 1986; Parish et al., 1990; Johnson and Caird, 1996; Hooper et al., 2007). Most of the ROI algorithms were not evaluated with Deaf users and are not real-time. Research into low frame rates for sign language are inconsistent in their conclusions, but there appears to be a sharp drop off in intelligibility at frame rates lower than 10 *frames per second* (fps).

MobileASL is built on top of the latest standard in video compression, H.264 (Wiegand et al., 2003). The H.264 encoder works by dividing a frame into 16×16 pixel *macroblocks*. It compares each macroblock to those sent in previous frames, looking for exact or close matches. The macroblock is then coded with the location of the match, the displacement, and whatever residual information is necessary. We use the Open Source x264 (Aimar et al., 2005; Merritt and Vanam, 2007) codec.

2. Design of the MobileASL System

The design of the MobileASL system is closely based on the needs and desires of users, and informed by a focus group and user studies.

2.1. Focus group

In our initial focus group, we find that users want a “smart” phone that has a front-side camera, a full keyboard, full email and instant messaging abilities, and a kick stand so that the phone can be placed on the table. Users also want to be able to use the phones to access video relay services, which allow communication between Deaf and hearing via sign language interpreters, and to chat with other users who have web cams or set top boxes. Based on these results, we choose to use HTC TyTN-II smart phones running Windows Mobile 6.1 (Qualcomm MSM7200, 400 MHz ARM processor, Li-polymer battery). The video size is QCIF (176×144). Figure 1 shows a phone running MobileASL. Our system is not currently able to handle calls to other devices, but we hope to add that functionality in the future.



Figure 1: MobileASL running on the HTC TyTN-II

2.2. Initial ROI and VFR evaluation

In several initial user studies, we investigate the feasibility of our ROI and VFR techniques. We find that videos with ROI are intelligible, up to a point; however, when too many bits are devoted to the face at the expense of the rest of the frame, it becomes distracting for users. For the variable frame rate, users evaluate conversational sign language videos that have (artificially created) lower frame rates during periods of not signing. We find that users dislike an entirely frozen frame for the not signing portions, but otherwise rate the quality similarly. As there is no large drop off in the perception of intelligibility, we use both methods in our system.

3. Sign language sensitive compression

To address the three main challenges of weak processing speed, limited bandwidth, and low battery life, we implement the following techniques for sign language sensitive video compression: optimal parameter selection for encoder optimization, dynamic skin-based ROI, and variable frame rate.

3.1. Optimal parameter selection

The H.264 encoder has many different parameters that are possible to tune to achieve the highest quality possible video at the lowest possible cost. For example, there are several different methods for searching the macroblocks for matching, with varying complexity. However, it is computationally infeasible to test all possible combinations of parameter settings for a given bit rate. Using a variation of the GBFOS (Chou et al., 1989) and ROPA (Kiang et al., 1992) algorithms, we jointly optimize H.264 encoder parameter settings for quality and complexity. We are able to search through many fewer encodings to arrive at the optimal selection.

3.2. Dynamic skin-based ROI

Given the parameter settings, H.264 will try to encode the frame at the highest possible quality for the bit rate. One way to increase intelligibility while maintaining the same bit rate is to shift the bits around, so that more are focused on the face and less are focused on the background. Using

a simple range query on the chrominance components, we determine the macroblocks that contain a majority of skin pixels, and encode these at a higher quality setting (allocating more bits to the important part of the frame). Since the encoder is constrained by the bit rate, the result is that the other macroblocks in the frame are encoded with fewer bits and correspondingly lower quality.

3.3. Variable frame rate

Sign language video is conversational and involves turn-taking, meaning that often when one person is signing, the other person is not. We aim to automatically recognize when a user is not signing and lower the frame rate from 10 fps to 1 fps. Since far fewer frames are encoded and transmitted, this results in a large power savings, allowing conversations to go on much longer. We obtain a power gain of 8% over the battery life of the phone, corresponding to an extra 23 minutes of talk time.

Automatic recognition on the phone is challenging for the same reasons as the overall system implementation. We must be able to perform the recognition in real-time while hopefully not adding to the complexity. To this end, we use a simple differencing method to distinguish signing frames from not signing frames. The sum of absolute differences of the luminance component is calculated between successive raw frames and compared to a previously determined threshold. This is temporally smoothed by applying a sliding window that takes the average vote over the window and classifies the frame accordingly. The average classification accuracy as measured on a frame-by-frame basis on videos taken with the phone camera is 76.6%.

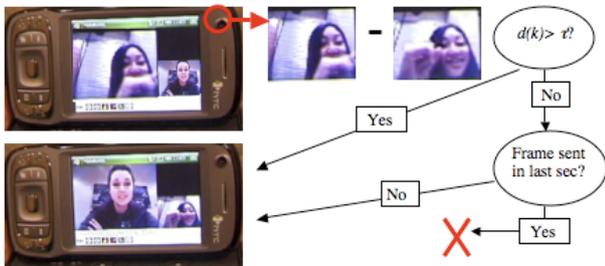


Figure 2: The architecture of the variable frame rate. Differences between frames are checked; if the user isn't signing, the frame is sent only to maintain 1 fps.

4. Evaluation

To validate our algorithms and test our working system, we conducted a user study with members of the signing Deaf community. Fifteen participants fluent in sign language took part in the study. For each conversation, participants sat on the same side of a table, separated by a screen, with a black background behind them (see Figure 3). Since we expect that Deaf people will use the phones in a variety of situations, we did not control for the relationship between participants. There were conversations between interpreters and native signers, between strangers and friends, and even between a married couple.

All combinations of three versions of ROI (no, low, and high) and two versions of VFR (off and on) were tested, for



Figure 3: Study setting. The participants sat on the same side of a table, with the phones in front of them.

a total of six different possible settings. After five minutes of unconstrained conversation, the participants filled out a subjective questionnaire about their experience. They then continued their conversation under different settings. The order in which the settings were evaluated differed between users. Both sides of the conversations were captured by a third video camera, in order to obtain objective measures, such as the number of times a user asked for a repetition.

We statistically analyzed both the subjective and objective results of the user study. For the subjective measures, we found statistically significant differences in the perception of the number of guesses and comprehension. Using a high level of ROI decreased the number of guesses and increased comprehension. ROI did not statistically significantly affect the objective measures, but VFR did. The users asked for repeats more often and had more conversational breakdowns when the VFR was on than when it was off. This is probably due to classification inaccuracy resulting in mistakenly lowering the frame rate when the person is actually signing. Despite these measurable difficulties with VFR, there was no statistically significant difference in subjective measures for VFR; in particular, the users' perceived desire for the technology was unaffected. We expect that VFR is a feature that users will choose to employ depending on their needs, for example, if they are going on a trip and want to preserve battery life.

5. Future directions

In the future, we will continue to improve MobileASL so that we may make it widely available. Our next step is to move out of the lab and into the field. We plan to give participants phones with MobileASL installed and have them use and comment on the technology over an extended period of time.

Technically speaking, several challenges remain. We can improve classification accuracy by using more advanced machine learning techniques on the phone. We found in our user study that often our algorithm misclassified finger spelling frames, since users slowed down during those periods. If our classifier recognized finger spelling in addition to signing and not signing, we could adjust the frame rate accordingly. We also want to investigate different methods for saving power on the phone, such as changing the spatial resolution during not signing periods instead of lowering the frame rate. Furthermore, there is a continual trade-off

in our system between the complexity of our algorithms, the speed at which we can encode, the intelligibility of the video, and the bit rate. We want to further explore jointly optimizing these conditions, ideally in real-time and as circumstances differ. For example, the encoder often struggles in noisy environments where there is a lot of background motion; in order to keep sending the frames in real time, we can reduce the quality, readjusting the parameters when circumstances improve.

The first question asked by users at the end of our study was always “when will this be available?” During the recruitment process, we received interested queries from all over the United States. Our ultimate goal is to make our technology widely available, so that Deaf people will have full access to today’s mobile telecommunication network.

6. References

- D. Agrafiotis, C. N. Canagarajah, D. R. Bull, M. Dye, H. Twyford, J. Kyle, and J. T. Chung-How. 2003. Optimized sign language video coding based on eye-tracking analysis. In *Visual Communications and Image Processing*, pages 1244–1252. SPIE.
- L. Aimar, L. Merritt, E. Petit, M. Chen, J. Clay, M. R., C. Heine, and A. Izvorski. 2005. x264 - a free h264/AVC encoder. <http://www.videolan.org/x264.html>.
- A. Cavender, R. E. Ladner, and E. A. Riskin. 2006. MobileASL: Intelligibility of sign language video as constrained by mobile phone technology. In *ASSETS '06: Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 71–78.
- N. Cherniavsky, A. C. Cavender, R. E. Ladner, and E. A. Riskin. 2007. Variable frame rate for low power mobile sign language communication. In *ASSETS '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*. ACM Press.
- N. Cherniavsky, R. E. Ladner, and E. A. Riskin. 2008. Activity detection in conversational sign language video for mobile telecommunication. In *Proceedings of the 8th international IEEE conference on Automatic Face and Gesture Recognition*. IEEE Computer Society.
- Neva Cherniavsky, Jaehong Chon, Jacob O. Wobbrock, Richard E. Ladner, and Eve A. Riskin. 2009. Activity analysis enabling real-time video communication on mobile phones for deaf users. In *UIST '09: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, October 4-7, 2009*, pages 79–88. ACM Press.
- Jaehong Chon, Neva Cherniavsky, Eve A. Riskin, and Richard E. Ladner. 2009. Enabling access through real-time sign language communication over cell phones. In *43rd Annual Asilomar Conference on Signals, Systems, and Computers*. IEEE Computer Society.
- Philip A. Chou, Tom D. Lookabaugh, and Robert M. Gray. 1989. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315.
- R. A. Foulds. 2006. Piecewise parametric interpolation for temporal compression of multijoint movement trajectories. *IEEE Transactions on information technology in biomedicine*, 10(1):199–206.
- N. Habili, C.-C. Lim, and A. Moini. 2004. Segmentation of the face and hands in sign language video sequences using color and motion cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(8):1086–1097.
- S. Hooper, C. Miller, S. Rose, and G. Veletsianos. 2007. The effects of digital video quality on learner comprehension in an American Sign Language assessment environment. *Sign Language Studies*, 8(1):42–58.
- B. F. Johnson and J. K. Caird. 1996. The effect of frame rate and video information redundancy on the perceptual learning of American Sign Language gestures. In *CHI '96: Conference companion on Human factors in computing systems*, pages 121–122. ACM Press.
- S. Z. Kiang, R. L. Baker, G. J. Sullivan, and C. Y. Chiu. 1992. Recursive optimal pruning with applications to tree structured vector quantizers. *IEEE Transactions on Image Processing*, 1(2):162–169.
- L. Merritt and R. Vanam. 2007. Improved rate control and motion estimation for H.264 encoder. In *ICIP '07: Proceedings of the 2007 IEEE International Conference on Image Processing*, volume 5, pages 309–312. IEEE Computer Society.
- D. H. Parish, G. Sperling, and M. S. Landy. 1990. Intelligent temporal subsampling of American Sign Language using event boundaries. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):282–294.
- D. M. Saxe and R. A. Foulds. 2002. Robust region of interest coding for improved sign language telecommunication. *IEEE Transactions on Information Technology in Biomedicine*, 6:310–316.
- R. Schumeyer, E. Heredia, and K. Barner. 1997. Region of Interest Priority Coding for Sign Language Videoconferencing. In *IEEE First Workshop on Multimedia Signal Processing*, pages 531–536. IEEE Computer Society.
- G. Sperling, M. Landy, Y. Cohen, and M. Pavel. 1986. Intelligible encoding of ASL image sequences at extremely low information rates. In *Papers from the second workshop Vol. 13 on Human and Machine Vision II*, pages 256–312, San Diego, CA, USA. Academic Press Professional, Inc.
- Rahul Vanam, Eve A. Riskin, and Richard E. Ladner. 2009. H.264/MPEG-4 AVC encoder parameter selection algorithms for complexity distortion tradeoff. In *DCC '09: Proceedings of Data Compression Conference*, pages 372–381. IEEE Computer Society.
- T. Wiegand, G. J. Sullivan, G. Bjntegaard, and A. Luthra. 2003. Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(7):560–576.
- W. W. Woelders, H. W. Frowein, J. Nielsen, P. Questa, and G. Sandini. 1997. New developments in low-bit rate videotelephony for people who are deaf. *Journal of Speech, Language, and Hearing Research*, 40:1425–1433.

Distributed System Architecture for Assisted Annotation of Video Corpora

Christophe COLLET, Matilde GONZALEZ, Fabien MILACHON

IRIT (UPS - CNRS UMR 5505)

Université Paul Sabatier, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9

{collet, gonzalez, milachon} AT irit DOT fr

Abstract

This paper present one component of Dicta-Sign, a three-year FP7 ICT project that aims to improve the state of web-based communication for Deaf people. A part of this project is the annotation of sign language corpora. To improve the annotation task in terms of reproducibility and time consuming, several plug-ins for sign language video processing are developed. The component presented in this paper aims to link several plug-ins to annotation software through the network. These plug-ins can be coded in different languages, operating systems and computers. For that, it uses the SOAP Web-service and a specific data-format in XML for the data exchange.

1. Introduction

Nowadays many researches focus on the analysis and recognition of sign language to understand, reproduce and translate to any other communication language (Ong and Ranganath, 2005). In computer science those researches concern the development of automatic treatments applied to sign language videos (Lefebvre-Albaret and Dalle, 2009; Theodorakis et al., 2009). The evaluation of their performances uses annotated corpora which is, in general, manually performed by linguists and computer scientists. Several Annotation Tools (AT) have been developed to achieve this task, e.g. Elan (Wittenburg et al., 2006), Anvil (Kipp, 2001), Ilex (Hanke, 2002; Hanke and Storz, 2008), Ancolin (Braffort et al., 2004), etc. For long video sequences, manual annotation becomes error prone, unreproducible and time-consuming. Moreover the quality of the results mainly depends on the annotator's knowledge. Automatic video processing together with the annotator's knowledge facilitate the task and considerably reduce the annotation time. That is why we propose a way to integrate those automatic treatments, here called Automatic Annotation Assistant (A^3), to the available AT.

From the annotator's point of view, adding automatic treatments must be easy to use, without adding complex A^3 calling or extra working. The annotator should be able to extract a part of a video and to use a previously defined annotation as input parameter of the A^3 . For example, the annotator is working in the Annotation tool window (fig. 1.a), any modification done is saved on the two tiers: AG1 and AG2. When the annotator calls an A^3 , e.g. movement pose detection, which needs two input parameters, then two additional tiers appear in the window (fig. 1.b). Filling in the two tiers could be done manually or using AG1 and/or AG2. Once the treatment has finished the result is displayed as a new tiers that the annotator can easily save or modify (fig. 1.c). This example shows how using automatic processing in this way can be easily performed.

The complexity of integrating the A^3 to the AT is not just about programming an efficient user friendly interface but also about making A^3 s and ATs to communicate with each other knowing that the programming environment used to developed them is not generally compatible. So, in this paper we propose a system architecture to allow the com-



Figure 1: Annotation Tool Example: (a) Normal environment, (b) A^3 call and (c) A^3 result.

munication between the A^3 s and the AT. We mainly focus on specifying communication protocols, data exchange and format.

This document presents the specification for this distributed system for assisted annotation of video corpus. First, we present a global view of the system. It consist of an overview of the architecture of the system proposed. Second, we illustrate the different communication between each sub-part of the system and the data format used to make them communicate. Finally, we present our choice about development software to use and about the security of the system.

2. Global view

The main problem about the introduction of A^3 s to existing ATs is the incompatibility of programming language, operative system and platform of development. Nevertheless it is not possible to restrict unique development conditions to easily use an A^3 to assist the annotation. Moreover treatments can be very complex and it would be preferable to develop them in a specific programming language or, even to execute them in adapted computers. That is why we proposed to overcome this problem by a Distributed System Architecture (DSA) where the A^3 s are hosted in different computers.

The communication and the data exchange are, then, done trough the network using a protocol and an exchange data format understandable by all the parts of the system. The data format has to be standardized so that the ATs and the A^3 s are able to process the data regardless where it

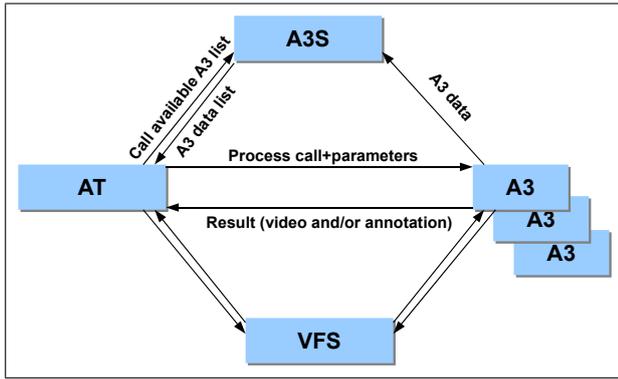


Figure 2: Distributed system architecture for assisted annotation of video corpus

comes from. Thus the A^3 's Application Programming Interface (API) has to use compatible parameters with the AT data structure. The data description standards proposed are XML and Annotation Graph (AG) structure (Bird and Liberman, 2001; Schmidt et al., 2008). The AG is a structure similar to the one used in the ATs, i.e. the hierarchy of named tiers (or levels or tracks...) with a list of possible values associated to each frame sequence. In this way any input parameter needed by the A^3 , can be filled in by the annotator with the help of the AT. In addition ATs are, generally, able to easily import/export AG structures. The AG is stored in a XML file which is extended to add the metadata concerning the desired A^3 processing and the video file.

The proposed DSA is illustrated in Figure 2. The principle is to consider the AT as a client and the A^3 s as remote servers to allow queries exchange. Since the number of available A^3 s and ATs can vary on time depending on new developments, another server called Automatic Annotation Assistant Supervisor (A^3S) is added to manage the information of the A^3 s at our disposal and to maintain an updated list of them. Thus at each time an A^3 is added it registers itself to the A^3S . Then when the AT requires an updated list of A^3 s it requests the A^3S server. Now the AT can directly communicate with the A^3 as long as the A^3 descriptor is known. In addition the need of exchanging video files between ATs and A^3 s leads to introduce a Video File Server (VFS) to share videos in a simple and fast way.

3. DSA data exchange

The AT allows annotators to easily define and execute various queries in a controlled manner. It interacts with all the parts of the system. Firstly, for the initialization process it queries the A^3S . Secondly, to process video it communicates to the respective A^3 . Finally, to add or to retrieve processed video files, it interacts with the VFS.

The A^3 communicates with the A^3S to register itself when it is added. All those interactions are illustrated in Figure 3

3.1. A^3 Registration

Each A^3 is considered as a unit implementing various processing functions for the annotation. To reference these

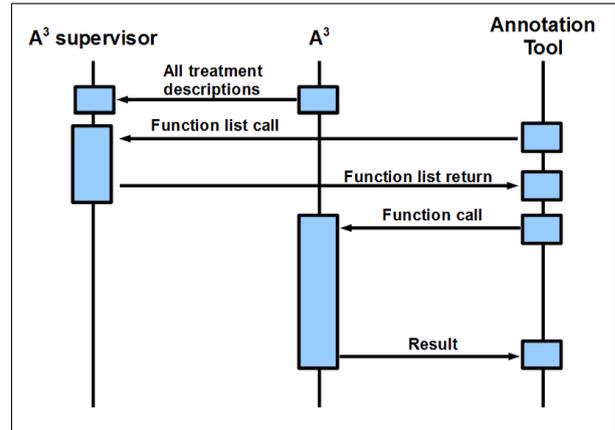


Figure 3: Simple Query Schedule

functions, each time an A^3 is added, it transmits its descriptor to the A^3S . The descriptor is a XML code containing API which has, among other information, a unique identifier (ID), the address (@), the port number (P) and the help text.

3.2. AT Initialization

The first query is automatically performed by the AT when it is loaded. This query is sent to the A^3S to ask for the list of available A^3 descriptors. The list can also be manually requested for updating at any moment. It does not require any parameter. In return, the A^3S sends the list of A^3 s and their descriptor. The AT can therefore decode the list and show to the annotator the available A^3 's functions descriptions.

3.3. A^3 calling

When the annotator selects a function, the parameters of this function are set up by filling in the AG provided by the A^3 descriptor. The minimum functionality that is expected from the AT, is an interactive editor for this AG. Two indispensable data elements are a list of videos and the ID of the process. The list of videos could correspond to different views for a same corpus. Once processing has been performed, it encapsulates the results, again in the form of an AG structure and sends them as a reply to the requesting AT.

4. XML Format

Previously, we defined that the API of each A^3 process uses a data format similar to the one used in AT, the AG. Due to the diversity of Annotation systems used, we need a simple global and compatible annotation graph system. That is why we decided to use an annotation graph format based on the one defined in Schmidt et al. (2008). The one we define is simple, easy to use and open. Thanks to this format, client can easily define frames and parameters for each frames, to use. In the server side, it has to read this AG to get the parameters needed to its process, execute it and finally put the result in the same AG. In order to simplify the processing of the parameters of the API and to get coherence, the video names, the process to call and every parameter are defined in this AG.

Finally, we have all annotation data, input and output, temporally described in an AG and encapsulated in one XML file. In addition this file contains the whole informations like API, option parameters, process descriptor and video location.

This AG encoded in XML format is described here step by step, through a simple API example.

```
<Metadata>
  <Parameter_in position="1">
    <Name value="Threshold_1"/>
    <Type value="Integer"/>
    <Default_value value="0,5"/>
    <Source value="A3_UPS"/>
  </Parameter_in>
  <Parameter_out position="2">
    <Name value="Coord_y_1"/>
    <Type value="Double"/>
    <Source value="A3_UPS"/>
  </Parameter_out>
```

Figure 4: XML data exchanged: Input and output parameters metadata

The figure 4 describes an input parameter and an output parameter in the API. Thanks to XML, this format can be easily parsed to get the different information about the input parameter, like its type and its default value. Each different parameter use its own `Parameters_in` or `Parameters_out` tag with a different position number.

```
<Process_Metadata>
  <Name value="Process1"/>
  <IP_address value="127.0.0.1"/>
  <Port value="8080"/>
  <Source value="A3_UPS"/>
  <Help>
    Here the HELP of the process.
  </Help>
</Process_Metadata>
```

Figure 5: XML data exchanged: Process metadata

```
<Video_Metadata>
  <Name value=""/>
  <IP_address value="127.0.0.1"/>
  <Port value="8008"/>
  <Source value="SFTP"/>
  <Login login="" pwd="" />
</Video_Metadata>
</Metadata>
```

Figure 6: XML data exchanged: Video metadata

Figure 5 and figure 6 describe respectively metadata about the process described by the API and inform about treated

video location. Most of the process metadata are about location of the process too. Moreover, it encapsulates the identification parameters (login and password) for the video server. If identification is needed to call process too, it can be easily added in the process metadata one the same way.

```
<Timeline id="A3_Timeline1">
  <Signal id="A3_Timeline1_Signal1"
    unit="frames" mimeType=""
    mimeType="" encoding=""
    xlink:href=""/>
</Timeline>

<AG timeline="A3_Timeline1" id="A3_AG1">
  <Anchor id="T0" offset="0" unit="frames"/>
  <Anchor id="T1" offset="1" unit="frames"/>
  <Annotation id="Annotation_1_T0"
    type="Parameter_in_1"
    start="T0" end="T0">
    <Feature name="Threshold_1"
      value="0,5"/>
    <Feature name="coord_x" value=""/>
  </Annotation>
</AG>
```

Figure 7: XML data exchanged: Annotation Graph (Time-Data)

Finally, figure 7 is a classical use of AGlib (Annotation Graph library) with definition of two time anchor and one AG in between. This AG contains, at the beginning, the minimal data: input parameters with default values and empty output results.

So, this format enables to represent all needed metadata. The user of the Annotation Tool has just to fill some parameters, like the video to use, and add each anchor and each annotation he needs. Afterward he will fill in those annotations with the desired input parameters and their values. When it is done, he sends this XML and the A³ will decode what it has to do. To send the result, it will automatically create result Annotation (and anchor if it need an anchor couple for each frame), fill their value, and send them.

5. Software development

We need a software library for network programming, which allows to develop this fairly simple architecture. Development constraints are that this system must be multi-platform and multi-language - including for the annotation software : RealBasic, C/C++ and Java - therefore we exclude proprietary libraries such as Java RMI or Twisted. Most of the time, the data to be transmitted are already in XML format, so a string can suffice. To achieve this kind of system two types of library are distinguished: the middleware - ICE (ZeroC, URL), CORBA (ObjectManagementGroup, URL) - and the Webservices - SOAP (W3C, URL), XML-RPC (XML-RPC, URL). The main difference between these two categories is that the first one, the middleware, is based on the use of objects and method calls on these objects, while the second one, web-services, is based

on the use of messages sent to URLs. It should be noted that each of these technologies meet our expectations, with different degrees of difficulty and complexity. We decided to use SOAP for the first specification of this architecture because it meets our needs in a simple way and it is totally open-source.

6. Security

In this system, we need security measures especially concerning the video corpus database. Indeed, those video files are not necessarily publicly accessible. To implement a sufficient security level, we propose two components: a secure transfer protocol, HTTPS, to transfer data by SOAP ; and a SFTP protocol for the transfer of video between the video file host and A³ or Annotation Tool.

7. Conclusion

In conclusion we propose a communication system architecture to easily add and call automatic treatments supporting annotation task in existing annotation tools.

Thanks to our specifications and the use of SOAP for software development, this architecture is multi-platform and multi-language (including RealBasic, C/C++ and Java) and the model used for data exchange is adaptable to many annotation formats. Furthermore, this model contains every needed information like location of the process to call, video to use, input and output parameters. The system is composed of four parts : the Annotation Tool (AT), the Automatic Annotation Assistants (A³), the A³ Supervisor (A³S) and a Video File Server (VFS). All communications between those entities are made through SOAP and are secured.

The programming of this system is underway during the project Dicta-Sign, and will enable to do evaluations of the contribution of automatic annotation process during annotation tasks

For future work we intend to enable asynchronous communication between AT and A³ in order to avoid waiting for the end of a long process (more than few seconds) and enable a deferred query for results. We also would like to enable time synchronized communication between the AT and interactive applications like the Signing avatar synthesizer (Kennaway et al., 2007) or Signing space annotation tool (Lenseigne and Dalle, 2005).

8. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

9. References

S. Bird and M. Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(Issues 1-2):23–60, January.

A. Braffort, A. Choisier, C. Collet, P. Dalle, F. Gianni, B. Lenseigne, and J. Segouat. 2004. Toward an annotation software for video of sign language, including image processing tools and signing space modelling. In *Proc.*

of 4th International Conference on Language Resources and Evaluation - LREC 2004, volume 1, pages 201–203, Lisbon, Portugal, May.

T. Hanke and J. Storz. 2008. ilex - a database tool for integrating sign language corpus linguistics and sign language lexicography. In *Proc. of 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages W25–64–W25–67, Marrakesh, May.

T. Hanke. 2002. ilex - a tool for sign language lexicography and corpus analysis. In *Proc. of 3rd International Conference on Language Resources and Evaluation, LREC 2002*, pages 923–926, Las Palmas de Gran Canaria, Spain.

J.R. Kennaway, J.R.W. Glauert, and Zwitterlood I. 2007. Providing signed content on the internet by synthesized animation. *ACM Transactions on Computer-Human Interaction*, 14(3):15/1–19, September.

M. Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proc. of 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370.

F. Lefebvre-Albaret and P. Dalle. 2009. Body posture estimation in a sign language video. In *Proc of The 8th International Gesture Workshop*, Feb.

B. Lenseigne and P. Dalle. 2005. Using signing space as a representation for sign language processing. In *Proc. of 6th International Gesture Workshop - GW 2005*, pages 25–36, Berder Island, France, 18-20 May. Springer-Verlag.

ObjectManagementGroup. URL. Corba documentation. <http://www.omg.org/technology/documents>.

S.C.W. Ong and S. Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 873–891.

T. Schmidt, S. Duncan, O. Ehmer, J. Hoyt, M. Kipp, D. Loehr, M. Magnusson, T. Rose, and H. Sloetjes. 2008. An exchange format for multimodal annotations. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, pages 207–221, Marrakech, Morocco, may. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.

S. Theodorakis, A. Katsamanis, and P. Maragos. 2009. Product-hmms for automatic sign language recognition. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 00, pages 1601–1604. IEEE Computer Society.

W3C. URL. Soap documentation. <http://www.w3.org/TR/soap/>.

P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. 2006. Elan: a professional framework for multimodality research. In *Proc. of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559.

XML-RPC. URL. Xml-rpc documentation. <http://www.xmlrpc.com/spec>.

ZeroC. URL. Ice documentation. <http://www.zeroc.com/download/Ice/3.4/Ice-3.4.0.pdf>.

Why are you raising your eyebrows?

Genny Conte¹, Mirko Santoro¹, Carlo Geraci², Anna Cardinaletti¹.

¹University Ca' Foscari-Venice, ²University of Milan-Bicocca.

Address: Carlo Geraci, Department of Psychology

Piazza dell'Ateneo Nuovo, 1

20126 – Milano, Italy

E-mail: carlo.geraci@unimib.it

Abstract

It is widely known that sign languages make an extensive use of non-manual markers (NMM) to transmit linguistic information. Some NMMs are specific to particular constructions (in several Sign Languages, furrowed eyebrows is mostly used to mark wh-questions, while headshake is used to mark negation), others may occur in several unrelated constructions (see eyebrow raising in American sign language). This study presents preliminary results of a quantitative investigation of the distribution of raised eyebrows (re-NMM) in Italian Sign Language (LIS). Re-NMM frequently occurs in spontaneous signing and is used to mark a variety of constructions; therefore re-NMM qualifies as a good candidate for a VARBRUL analysis. In particular, re-NMM may mark 8 different constructions in LIS: yes/no-questions, topics, if-clauses, correlative clauses, focus, contrastive focus, subordinate clauses, and the signer's attitude. Data come from a corpus of LIS and have been analyzed with the ELAN software. Results show an even distribution across the sample for most of the uses of re-NMM. Only two functions turned out to be significantly different: the use of re-NMM as a focus marker and the use of re-NMM as an attitude marker, which are sensitive to age.

1. Introduction

One of the most interesting properties of sign languages (SLs) is the use of non-manual components to transmit linguistic information. At a first glance, non-manual markers (NMMs) can be thought to have a similar role to that played by prosody in spoken languages. While this is certainly true (see for instance Nespor and Sandler 1999), it is also clear that NMMs are much more than that. Indeed, they represent a pervasive aspect of SLs. All levels of linguistic analysis are affected by the presence of NMMs: they are productively used to mark specific lexical items, and in some cases they also mark phonological contrast (see Franchi, 2004, for some examples from Italian Sign Language, LIS). They are used as adverbial markers (see Neidle et al. 2000, for some examples from American sign language, ASL). They can also be used as markers of discourse features like the signer's attitude and more generally as affective markers. NMMs have an impact also in the domain of semantics. For instance, in some varieties of LIS, the position of the shoulder is used to mark the event time (Zucchi, 2009). However, the most intriguing use of NMMs is in the domain of syntax, where NMMs play a crucial role in determining several syntactic functions and constructions such as overt agreement (Neidle et al. 2000), negation (Neidle et al. 2000, Geraci, 2006 and Pfau & Quer, 2007 among others), wh-questions (Cecchetto, Geraci & Zucchi 2009), etc.

Several independent articulators can be used to produce NMMs and, most importantly, they can act simultaneously so that a certain degree of overlapping is generally allowed. For instance, (raised or lowered) eyebrow positioning may co-occur with head-tilt, eye gaze, and some specific body postures. As discussed in Wilbur (2000), the main function of NMMs is to single out specific linguistic domains. Depending on the

articulator(s), this can be done either by signalling domain boundaries (as in the case of eye blinking or head nods), or by spreading the marker over the whole domain (as in the case of headshake or eyebrow positioning). In the former case, NMMs are used as edge-makers, while in the latter case they are used as scope markers. Within the class of scope markers, raised eyebrows pose a particular challenge. Indeed, while headshake and, to a certain extent, furrowed eyebrows can be argued to mark specific constructions (negation and wh-questions, respectively), raised eyebrows are found to occur with several and apparently unrelated constructions (for ASL, see Wilbur 2000). The aim of this study is twofold: on the one hand, we analyze the distribution of the raised eyebrow NMM (re-NMM) in LIS; on the other hand, we investigate whether non-linguistic factors may have a role in such distribution. In particular, it is likely that social factors may affect the use of re-NMM and the variety of constructions in which it occurs. This is accomplished by presenting preliminary results of a quantitative analysis of the distribution of re-NMM in a corpus of LIS data (Geraci et al. 2010).

2. The re-NMM variable

To our knowledge, there is no systematic investigation of re-NMM in LIS. However, the presence of this marker is observed in many studies, and it is associated with a variety of constructions. In particular, re-NMM is associated with:

- Yes/no questions (Cecchetto, Geraci & Zucchi 2006),
- If clauses (Barattieri, 2006),
- (Cor-)relative clauses (Cecchetto, Geraci & Zucchi 2006, Geraci, 2007, Branchini & Donati, 2009),

- Topicalized elements (Geraci, 2006 and Geraci, Cecchetto & Zucchi, 2008 and Bertone, 2009),
- Subordinate and complement clauses (Geraci, 2007, and Geraci, Cecchetto & Zucchi 2008).

Other previously unnoticed uses of re-NMM emerged in this study are:

- Broad focus,
- Contrastive focus,
- Emphatic discourse attitude.

Of course, as it happens with other NMMs, re-NMM is not the exclusive marker for the above-mentioned constructions. Other non-manual components may co-occur with it, or it can also be the case that re-NMM is only one of the possible means to mark the construction. Be as it is, such variation of uses is likely to be influenced not only by purely linguistic factors, but also by non-linguistic factors (such as age and gender). Furthermore, given its highly frequent distribution, re-NMM nicely qualifies as a candidate for a variation analysis with standard sociolinguistic techniques (Bayley 2002).

3. Data collection

The data from this study comes from a corpus of LIS which is under construction as part of a national research project on sociolinguistic variation in LIS (see Geraci et al. 2010). The corpus includes data from signers of three age groups (18-30, 31-54, over 55) recruited in 10 cities distributed across the country and consists of various kinds of texts, namely free conversation (45 minutes), elicited conversation (about 5-10 minutes), individual narration (10 minutes), and a picture-naming task (42 items). For this study, we analyzed the narrative production of 16 signers from the city of Torino. Six signers were in the group of old signers, while the middle and young groups consisted in five signers each. All participants agreed in being recorded. In order to avoid the situation of a signer sitting right in front of the camera and to reduce the potential negative effects of recording, signers were asked to sign to a Deaf addressee from the same local Deaf community. The camera was placed right behind the addressee, so that a frontal view of the narrator was provided. Signers were asked to tell stories about their life experience, nevertheless they were free to change topic at their pleasure.

4. Methodology

Data were analyzed by using the ELAN software (Johnston & Crasborn, 2006). The annotations were made by a LIS interpreter, enrolled as a second year MA student at the Università Ca' Foscari-Venezia, and were crosschecked by two native signers of LIS. For this study four tiers were employed, as shown in figure 1.

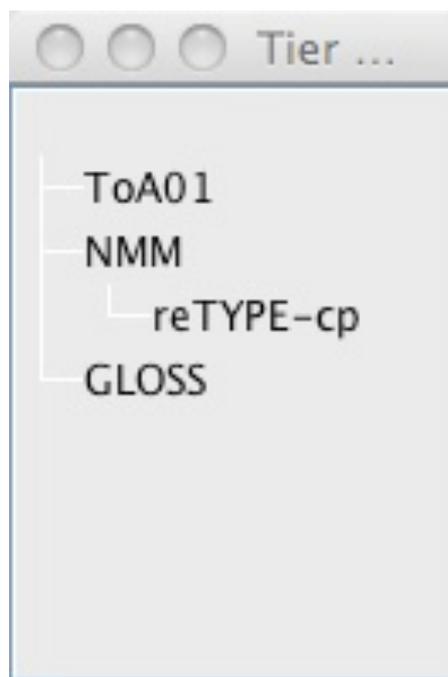


Figure 1: Tiers used for the re-NMM study

The main tier (ToA01, namely, old signer from Torino) includes the annotations of the utterance in which a re-NMM occurred. The GLOSS tier includes the sign-by-sign annotation of the utterance, while the NMM tier marks the spreading of eyebrows raising. Finally, the re-Type tier specifies which function is associated to that raising. Since the number of functions is limited, a controlled vocabulary has been created with 8 possible functions for the re-NMM: y/n question, if-clauses, (cor-)relatives, topic, subordination, focus, contrastive focus, and attitude. The procedure adopted for the annotation involved four steps: First, every occurrence of eyebrow raising was simply marked (NMM tier). Second, the annotations for the utterance were inserted (main tier). Third, the function of the re-NMM was selected (re-Type tier). Fourth, the gloss for each sign included in the utterance was provided (GLOSS tier). Figure 2 illustrates the ELAN workspace for this study.

5. Results

A total of 410 instances of re-NMM have been coded. The overall distribution for each function of re-NMM is given in table 1. Independently from the linguistic functions, old signers tend to use re-NMM (44.1 %) more than signers of the middle (30%) and young (25.1%) groups, and male signers (57.6%) tend to use re-NMM more than female signers (42.4%). Furthermore re-NMM is mostly used to mark broad focus (34.4%) and topic (26.8%). Apart from broad focus and attitude, the remaining functions of re-NMM are equally distributed across the factors in both factor groups (Age and Gender), as can be seen from the percentages reported in table 1.

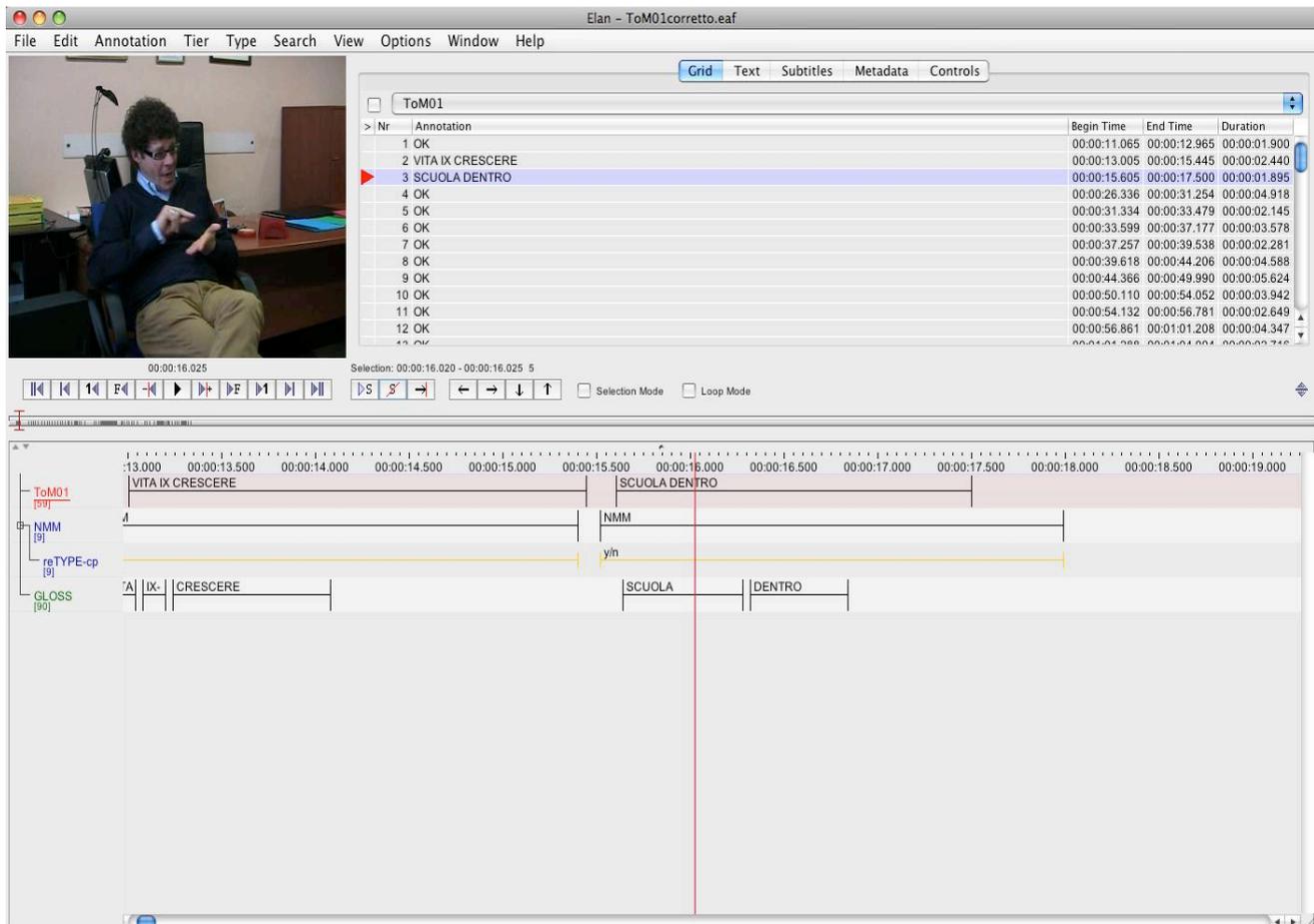


Figure 2: Workspace for the re-NMM study.

	Function	t	f	c	i	s	a	y	r	Total	%
Age	o	46	63	7	7	7	38	9	4	181	44.1
	%	25.4	34.8	3.9	3.9	3.9	21.0	5.0	2.2		
	m	36	34	9	6	12	12	13	4	126	30.7
	%	28.6	27.0	7.1	4.8	9.5	9.5	10.3	3.2		
	y	28	44	7	4	4	9	5	2	103	25.1
	%	27.2	42.7	6.8	3.9	3.9	8.7	4.9	1.9		
Gender	M	60	76	14	9	11	43	15	8	236	57.6
	%	25.4	32.2	5.9	3.8	4.7	18.2	6.4	3.4		
	F	50	65	9	8	12	16	12	2	174	42.4
	%	28.7	37.4	5.2	4.6	6.9	9.2	6.9	1.1		
	Total	110	141	23	17	23	59	27	10	410	
	%	26.8	34.4	5.6	4.1	5.6	14.4	6.6	2.4		

Table 1: Distribution of re-NMM functions by Factor Group. Functions: t = topic, f = broad focus, c = contrastive focus, i = if-clause, s = subordinate, a = attitude marker, y = y/n question, r = relative clause. Age: o = old signers' group, m = middle signers' group, y = young signers' group; Gender: M = male signers, F = female signers.

Indeed, only Age showed a significant effect in two of the eight VARBRUL analyses, performed with broad focus and attitude defined as the application value. Results for this factor group are shown in table 2. We have included the input value for each run, an overall measure of the tendency of signers to choose the application value and the chi-square per cell, a measure of the goodness of fit.

Factor	Broad Focus		Attitude	
	Weight	%	Weight	%
Old	.507	34.8	.632	21
Middle	.416	27.0	.405	9.5
Young	.590	42.7	.383	8.7
Input	.342	34.4	.134	14.4

Table 2: Functions of re-NMM by Age. Note: Broad Focus, $\chi^2/\text{cell} = 0.0660$; Attitude, $\chi^2/\text{cell} = 0.0764$.

On the one hand, the use of re-NMM to mark broad focus is favored by young signers ($p = .590$) and disfavored by middle signers ($p = .416$), while old signers neither favor nor disfavor the use of re-NMM to mark broad focus. On the other hand, the use of re-NMM as an attitude marker is favored by old signers ($p = .632$) and clearly disfavored both by signers from the middle ($p = .405$) and young ($p = .383$) groups.

6. Discussion

Eyebrows raising is a fundamental component of the grammar of sign languages. In LIS, as in ASL, this non-manual marker is widely used in several constructions. In particular, re-NMM is used to mark eight different linguistic functions. Interestingly, the data reported here show a significant effect of age in the use of re-NMM. In particular, young signers use re-NMM to mark broad focus more often than other age groups, and older signers tend to use re-NMM as an attitude marker while middle and young signers disfavor the use of re-NMM for this function. Both these effects can be interpreted as a diachronic tendency toward the use of re-NMM with a fine-grained linguistic function. Of course, more research and more data from other types of texts and other cities are needed to confirm this hypothesis and to evaluate how consistent our findings are with respect to the varieties of LIS signed in other cities.

References

Barattieri, C. (2006). *Il periodo ipotetico in LIS*. Siena: University of Siena master thesis.

Bayley, R. (2002). The quantitative paradigm. In J. K. Chambers, P. Trudgill, N. Schilling-Estes (Eds.), *The handbook of language variation and change*. Malden, MA: Blackwell Publishing, pp. 117--141.

Bertone, L. (2009). The syntax of noun modification in

Italian Sign Language (LIS). In *University of Venice Working Papers in Linguistic*, 19, pp 7--28.

Branchini, C., & Donati, C. (2009). Relatively different: Italian Sign Language relative clauses in a typological perspective. In A. Liptak (ed.), *Correlatives cross-linguistically*. Amsterdam: John Benjamins.

Cecchetto, C., Geraci C., & Zucchi, S. (2006). Strategies of relativization in LIS. *Natural Language and Linguistic Theory*, 24, pp. 945--975.

Cecchetto, C., Geraci, C., & Zucchi, S. (2009). Another way to mark syntactic dependencies: the case for right-peripheral specifiers in sign languages. *Language* 85(2), pp. 278--320.

Franchi, M.L. (2004). Componenti non manuali. In V. Volterra (Ed.), *La lingua dei segni italiana. La comunicazione visivo-gestuale dei sordi*. Bologna: Il Mulino, pp. 159--177.

Geraci, C. (2006). Negation in LIS (Italian Sign Language). In L. Bateman, C. Ussery (Eds.), *Proceedings of NELS 35*. Amherst, MA: GLSA, pp. 217--229.

Geraci, C. (2007). Comparative correlatives in Italian Sign Language. *TAL*, 48 (3), pp. 55--92.

Geraci, C., Cecchetto, C. & Zucchi, S. (2008). Sentential complementation in Italian Sign Language. In M. Grosvald & D. Soares, *Proceedings of the 38 Western Conference on Linguistics: WECOL 2008*, pp. 46--58

Geraci et al. (2010). Building a corpus for Italian Sign Language. Methodological issues and some preliminary results. In *Proceedings of LREC 2010*.

Johnston, T., Crasborn, O. (2006). The use of ELAN annotation software in the creation of signed language corpora. In *Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, MI.

Neidle, C., Kegl, J., Macloughlin, D., Bahan, B., & Lee, R.G. (2000). *The syntax of American Sign Language*. Cambridge, MA: MIT Press.

Nespor, M., & Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech*, 42(2-3), pp. 143--176.

Pfau, Roland and Josep Quer (2007). On the syntax of negation and modals in Catalan Sign Language and German Sign Language (LSC). In P. Perniss, R. Pfau, & M. Steinbach (Eds.), *Visible variation: Cross-linguistic studies on sign language structure*. Berlin: Mouton de Gruyter, pp. 129--161.

Wilbur R.B. (2000). Phonological and prosodic layering of nonmanuals in American Sign Language. In H. Lane & K. Emmorey (Eds.), *The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima*, Hillsdale, NJ: Lawrence Erlbaum, pp. 213--241.

Zucchi, S. (2009). Along the time line: tense and time adverbs in Italian Sign Language. *Natural language semantics*, 17(2), pp. 99--139.

Sign Language Recognition using Linguistically Derived Sub-Units

Helen Cooper, Richard Bowden

University Of Surrey
Guildford, UK.

H.M.Cooper@Surrey.ac.uk, R.Bowden@Surrey.ac.uk

Abstract

This work proposes to learn linguistically-derived sub-unit classifiers for sign language. The responses of these classifiers can be combined by Markov models, producing efficient sign-level recognition. Tracking is used to create vectors of hand positions per frame as inputs for sub-unit classifiers learnt using AdaBoost. Grid-like classifiers are built around specific elements of the tracking vector to model the placement of the hands. Comparative classifiers encode the positional relationship between the hands. Finally, binary-pattern classifiers are applied over the tracking vectors of multiple frames to describe the motion of the hands. Results for the sub-unit classifiers in isolation are presented, reaching averages over 90%. Using a simple Markov model to combine the sub-unit classifiers allows sign level classification giving an average of 63%, over a 164 sign lexicon, with no grammatical constraints.

1. Introduction

Sign Language Recognition (SLR) has many parallels to speech recognition, the idea which has been seized by many is that of combining sub-units into word level classifiers. Doing this has several advantages; it allows the lexicon to be increased in a manageable manner. It removes much of the temporal variance between repetitions of the same sign. It enables linguistics to be used, to add priors to the sub-unit combinations and it could feasibly lead to classification of unseen signs based on their component parts and a dictionary. For these last two advantages to be realised, the sub-unit classifiers need to be derived from the linguistic domain.

Previous systems using tracking-based, sub-unit classifiers, have tended to either hard code basic sub-units (Kadir et al., 2004) or used data driven approaches (Han et al., 2009; Yin et al., 2009). While both these techniques can give good sign level results, they bear little relation to the linguistics of sign language. Instead, the sub-unit classifiers proposed in this paper are learnt from data, annotated at the sub-unit level, using the same notation as that in the British Sign Language (BSL) Dictionary (British Deaf Association, 1992).

In this work, first the signer is tracked, then sub-unit, classifiers are learnt using boosting. Specifically, sub-units relating to Position (*Tab*), Hand Arrangement (*Ha*) and Movement (*Sig*) are covered. These classifier responses are also shown in combination with a Markov chain Look Up Table (LUT) to perform basic classification at the sign level. The details of these classifiers are shown in the following sections.

2. Method

Tracking results are obtained using Buehler *et al.*'s tracker which does not require coloured gloves, whilst still giving accurate results, on natural sign from TV broadcasts it achieves >80% (Buehler et al., 2008). The tracking system gives boxes bounding the hands, lower arms and upper arms. The different sub-unit types are catered for by different weak classifier concepts; *Tab* requires information about positioning, *Ha* about the relationship between

the hands and *Sig* about the temporal changes in hand positions, often relative to each other. Each of the different weak classifier types are combined using AdaBoost (Freund and Schapire, 1995) to create a classifier for each sub-unit present in the training set.

2.1. Tab Classifiers

Classifiers concentrating on *Tab* sub-units are concerned with spatial features, describing the location of the hands in relation to the signer. The bounding boxes of the hands are given by the tracking and the position of the face can be found using the Viola Jones face detector (Viola and Jones, 2001). Classifiers can then be built which consider relational distances. Each classifier operates on an x or y feature, i , within the tracking vector, \mathbf{o} , comparing it to an upper and lower limit, \mathcal{T}_U and \mathcal{T}_L respectively. If the value falls within this range, then the classifier fires. The upper and lower limits are individual to each classifier and calculated relative to the size (f) and position (f_{xy}) of the signer's face, see Equation 1.

$$\begin{aligned} \mathcal{T}_L &= f_{xy} + nf \\ \mathcal{T}_U &= \mathcal{T}_L + sf \\ n &\in \{-3, -2.9, -2.8 \dots 3\} \\ s &\in \{0.1, 0.2, 0.3 \dots 1\} \\ R_{wc} &= \begin{cases} 1 & \text{if } \mathcal{T}_L < \mathbf{o}_i \leq \mathcal{T}_U \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (1)$$

Classifiers can work on the x or y co-ordinates of either the dominant or non-dominant hand. Each classifier covers a strip of a given constant width, either in the x or y plane. Boosting is used to combine these weak classifier strips, to create areas relative to the signer as shown in Figure 1. The strips are shown by increasing the luminosity of the pixels. When many weak classifiers overlap, the area turns white. As can be seen, the white areas coincide with the area being learnt, *i.e.* Figure 1(a) 'face' and Figure 1(b) 'upper arm'.

2.2. Sig Classifiers

Sig sub-units describe the motion of a sign and require classifiers which encode temporal information. The tracking

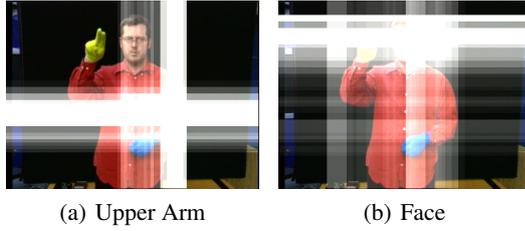


Figure 1: Examples of tracked *Tab* classifiers for the areas ‘upper arm’ and ‘face’. Boosting combines strips in the x and y planes to show where the hand is expected to be for each *Tab* label. The lighter the area in the picture the more strips are overlaying it.

provides a frame by frame set of co-ordinates for the hands so motions can be described by changes in these values. The sub-units from BSL linguistics do not encode magnitude information. Therefore the classifiers used to describe them need to encode non-magnitude dependant information. If the values from the tracking are concatenated temporally into 2D vectors, then it is possible to examine individual components across time. In this way, a weak classifier can look for changes in, for example, the x co-ordinate of the dominant hand. This would encode left and right motion of the dominant hand. Component values can either increase, decrease or remain the same, from one frame to the next. If an increase is described as a 1 and a decrease or ‘no change’ is described as a 0 then a Binary Pattern (BP) can be used to encode a series of increases/decreases. A temporal vector is said to match the given BP if every ‘1’ accompanies an increase between concurrent frames and every ‘0’ a decrease/‘no change’. This is shown in Equation 2 where $\mathbf{O}_{i,t}$ is the value of the component, \mathbf{o}_i , at time t and \mathbf{bp}_t is the value of the BP at frame t . See Figure 2 for an example where feature vector A makes the weak classifier fire, whereas feature vector B fails, due to the ringed gradients being incompatible.

$$R_{wc} = \left| \max_{\forall t} (BP(\mathbf{O}_{i,t})) - 1 \right|$$

$$BP(\mathbf{O}_{i,t}) = \mathbf{bp}_t - d(\mathbf{O}_{i,t}, \mathbf{O}_{i,t+1})$$

$$d(\mathbf{O}_{i,t}, \mathbf{O}_{i,t+1}) = \begin{cases} 0 & \text{if } \mathbf{O}_{i,t} \leq \mathbf{O}_{i,t+1} \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Discarding all magnitude information would mean that salient information might be removed. To retain this information, boosting is given the option of using additive classifiers as well. These look at the average magnitude of a component over time. The weak classifiers are created by applying a threshold, \mathcal{T}_{wc} , to the summation of a given component, over several frames. This threshold is optimised across the training data during the boosting phase. For an additive classifier of size T , over component \mathbf{o}_i , the response of the classifier, R_{wc} , can be described as in Equation 3.

$$R_{wc} = \begin{cases} 1 & \text{if } \mathcal{T}_{wc} \leq \sum_{t=0}^T \mathbf{O}_{i,t} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

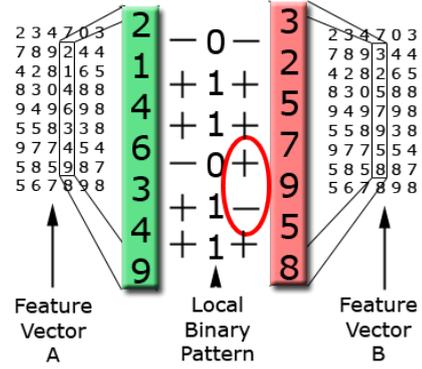


Figure 2: An example of a BP being used to classify two examples. A comparison is made between the elements of the weak classifiers BP and the temporal vector of the component being assessed. If every ‘1’ in the BP aligns with an increase in the component and every ‘0’ aligns with a decrease or ‘no change’ then the component vector is said to match (*e.g.* case A). However if there are inconsistencies as ringed in case B then the weak classifier will not fire.

Boosting is given all possible combinations of BPs, acting on each of the possible tracking components. The BPs are limited in size to being between 2 and 5 changes (3 - 6 frames) long. The additive features are also applied to all the possible components, but the lengths permitted are between 1 and 26 frames. Both sets of weak classifiers can be temporally offset from the beginning of an example, by any distance up to the maximum distance of 26 frames.

2.3. *Ha* Classifiers

Ha sub-units explain the hand arrangement present in a sign, *e.g.* which hand is higher or whether they are inter-linked. Using the tracked positions on each frame, the x and y values of all points can be compared. This can be done using a magnitude comparison, as illustrated in Equation 4 where $\mathbf{O}_{i,t}$ is the first component and $\mathbf{O}_{j,t}$ is the second, both on frame t . Though this does not encode any information about the magnitude of the difference required for the weak classifier to fire. Alternatively, for each point-comparison, 11 weak classifiers are built. Each requiring a different magnitude difference to fire. The difference magnitude, \mathcal{T}_{wc} , is selected from a set of 0 to 50 pixels in 5 pixel steps as shown in Equation 5. This selection of thresholds gives $(36!/(34!*2!)) * 11 = 6930$ possible weak classifiers.

$$R_{wc} = \begin{cases} 1 & \text{if } (\mathbf{O}_{i,t} \leq \mathbf{O}_{j,t}) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$R_{wc} = \begin{cases} 1 & \text{if } \mathcal{T}_{wc} \leq (\mathbf{O}_{i,t} - \mathbf{O}_{j,t}) \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{T}_{wc} = 0, 5, 10 \dots 50 \quad (5)$$

3. Data Set

This work uses the same 164 sign data set as Kadir *et al.* (Kadir *et al.*, 2004) but with extra annotation at the sub-unit level. 7410 *Tab* examples, 322 *Ha* examples and 578

Label	<i>Ha</i>	<i>HaM</i>
left_up	82.14%	87.21%
right_up	67.83%	93.88%
side_by_side	91.01%	93.85%
contact	81.45%	87.13%
left_nearer	91.67%	100.00%
right_nearer	96.43%	87.50%
interlink	73.68%	96.55%
Mean	83.46%	92.30%
Std Dev	5.13pp	10.31pp

Table 1: Results of *Ha* and *HaM* tracking based classifiers

Sig were hand labelled for training viseme classifiers. The data set consists of 1640 sign examples. Signs were chosen randomly rather than picking specific examples which are known to be easy to separate. *Tab* sub-units are static and happen on a single frame with multiple frames per sign. As such, the example counts are higher than those for *Sig* which are movement visemes and happen across multiple frames. *Ha* visemes are also static, however, they change more quickly within a sign than *Tab* visemes. As a result, there are often only one or two frames per sign which contain the *Ha* value given by the BSL Dictionary.

4. Sub-Unit Results

For the tracked classifiers, six different types of classifiers were tested for the three different sub-unit types. For *Ha* sub-units, there are two possible classifiers; those which make a binary comparison on the x and y positions of the hands, and those described in more detail in Section 2.3. where the magnitude of the difference is taken into account. For *Tab*, the two classifiers tested are based on the labelling, the first uses the labels independently, the second implements the hierarchical structures described in Section 2.1. The *Sig* sub-unit classifiers were tested with both the standard labels and the revised component labels.

Classifiers are trained on sub-units from four out of ten available signs, then tested on the sub-units from the remaining six. The results shown are taken from the diagonals of confusion matrices across each sub-unit type.

4.1. *Ha* Classifiers

First is the comparison between the results of the binary comparison *Ha* classifiers and the comparators which take the magnitude into account shown in Table 1. The former manage a good response with a mean true-positive rate of 83.46% achieving a maximum 96.43%. The classifiers which include magnitude manage better on all labels but one, with a true-positive mean of 92.30%, 9pp better than the previous results. The magnitude comparators also result in a more consistent classifier with a Standard Deviation (Std Dev) half that of the binary comparison classifiers.

4.2. *Tab* Classifiers

Next, the tracked *Tab* classifiers are examined with the original labels, see Table 2, the mean true positive classification rate is poor, achieving only 46.95% with some classifiers getting 0%. Notably where it fails to distinguish

between ‘upper_arm’ and ‘lower_arm’. Moving to the hierarchical label system, the first thing to note is that confusions are only considered between labels of the same level (e.g. ‘face’ is compared to ‘arm’ but not to ‘face_lower’ or ‘arm_upper’). This is because the data for some of the lower levels is used as positive training data for the higher labels, so a direct comparison cannot be made with the non-hierarchical labels due to the changes in the way the confusion matrices need to be constructed. However, when using these labels, in the confusion matrix, the mean true-positive rate is 79.84%, 33pp higher than the non-hierarchical version. There is also a reduction of 10pp in the Std Dev suggesting that this again gives a more consistent classifier.

Label	<i>Tab</i>	<i>TabH</i>
arm		97%
chest	80%	35%
face	47%	95%
arm_low	85%	71%
arm_up	0%	54%
chest_right	0%	88%
chest_up	75%	97%
face_low	53%	78%
face_side		75%
face_up	71%	81%
chest_up_shoulder		91%
face_low_mouth	30%	59%
face_low_nose	39%	83%
face_low_underchin	72%	95%
face_side_cheek	30%	67%
face_side_ear	30%	81%
face_up_eyes	25%	75%
chest_up_shoulder_right	69%	98%
Mean	46.95%	79.84%
Std Dev	27.90pp	17.73pp

Table 2: Results of *Tab*tracking based classifiers.

4.3. *Sig* Classifiers

The two versions of tracked *Sig* classifiers, like the previous tracked *Tab* classifiers, are based solely on a change in the way the training labels are used. The difference between the *Sig* classifiers and the other sub-unit classifiers, is that the *Sig* classifiers are boosted across more than one frame, so the training data is used not only to create the classifiers but also to choose the length of the chosen strong classifier. Confusion matrices are calculated for each possible length over the training data. Table 3 shows the results from the training and testing. *Sig* classifiers (using the original labels) give a training true-positive rate of 62%, which is substantially higher than the test average of 48% achieved when using the training derived lengths.

The outcome is similar when examining the results for the new component based labels *SigC*. The best training lengths give an average of 79% which is an increase of 17pp over the non component based training system. This is reflected in the results when using the training lengths on the test data, where a 53% level is attained, a 5pp increase on the previous result.

Label	SigC Train max		SigC Test	
	SigC Train max	SigC Test	Sig Train max	Sig Test
B_apart	98%	74%	97%	81%
B_circ_tog_down_alt	69%	41%	0%	0%
B_circ_tog_tow	82%	49%	0%	0%
B_down	63%	53%	78%	67%
B_tog	82%	52%	88%	66%
B_tow_away_alt	92%	45%	94%	44%
B_up	91%	71%	80%	72%
B_up_down	100%	93%	93%	87%
B_up_down_alt	100%	97%	0%	0%
D_away	58%	36%	75%	46%
D_away_down	67%	28%	0%	0%
D_circ_left_down	83%	84%	100%	95%
D_circ_left_tow	100%	98%	100%	98%
D_down	44%	40%	77%	74%
D_down_away	46%	20%	0%	0%
D_left	90%	48%	63%	61%
D_left_right	93%	51%	77%	33%
D_right	48%	27%	33%	28%
D_tap	67%	24%	44%	39%
D_tow	87%	26%	100%	51%
D_tow_away	91%	37%	76%	48%
D_wrist_tow_away	81%	58%	64%	34%
D_up	88%	74%	96%	90%
Mean	79%	53%	62%	48%
Std Dev	18%	24%	38%	33%

Table 3: Results of *Sig* classifiers and *SigC* classifiers using component based labels. The first column shows the maximum training classification achieved, the second shows the rate when using the length, found via training, on the test data.

5. Sign Level Results

For completeness, basic sign level results are shown using the same Markov Model as that in (Kadir et al., 2004) The second stage classifier is trained on the previously used four training examples plus one other, giving five training examples per sign. Shown in Table 4 as the results of combining the various sub-unit classifiers with the Markov model. The best results are gained using the magnitude comparisons for *Ha*, the hierarchical representation of *Tab* and the basic *Sig* classifiers, getting 63%.

6. Conclusions

Tests were conducted using boosting to learn three types of linguistic sub-unit, which are then combined with a simple second stage classifier to learn word level signs. By basing the sub-units on the linguistic taxonomy there is greater scope for using data and priors from the linguistic domain as well as using the sub-unit classifiers to aid in data annotation. However, this data set is few in repetitions, with only 4 per sign for training the viseme level classifiers. This means that there are not always enough examples to fully separate each viseme type and more information than just

Combination	<i>Ha</i>	<i>HaM</i>	<i>HaM</i>	<i>HaM</i>
	<i>TabH</i>	<i>Tab</i>	<i>TabH</i>	<i>TabH</i>
	<i>Sig</i>	<i>Sig</i>	<i>SigC</i>	<i>Sig</i>
Mean	35.7%	60.6%	55.5%	63.0%
Minimum	33.9%	57.7%	52.7%	61.2%
Maximum	36.6%	62.4%	57.1%	65.1%
Std Dev	0.8	1.6	1.4	1.3

Table 4: Classification performance using sub-unit level classifiers, combined together by a basic Markov Model LUT, trained on five examples. *Ha* uses binary comparisons between values, whereas *HaM* uses the magnitude of the difference between values. *Tab* does not use the hierarchical structure of this sub-unit class, *TabH* includes this structure. *SigC* uses the component based labels whereas *Sig* use the standard labels.

the viseme might be encoded by the classifier. It is also lacking in the number of signs it contains, having only 164 signs, which is insufficient to fully represent all the visemes for which classifiers should be learnt. However, there is currently no other publicly-available data set, which has sub-unit labelling at the temporal level, with which to better train the classifiers. It is for this reason that future work should investigate other sources of data whilst continuing to use a sub-sign representation allowing large lexicons to be tackled effectively.

7. References

- British Deaf Association. 1992. *Dictionary of British Sign Language/English*. Faber and Faber.
- P. Buehler, M. Everingham, D. P. Huttenlocher, and A Zisserman. 2008. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Procs. of BMVC*, pages 1105 – 1114, Leeds, UK, September 1 – 4.
- Yoav Freund and Robert E Schapire. 1995. A decision-theoretic generalization of on-line learning and an application to boosting. In *Procs. of European Conference on Computational Learning Theory*, pages 23 – 37, Barcelona, Spain, March 13 – 15. Springer-Verlag.
- J.W. Han, G. Awad, and A Sutherland. 2009. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pattern Recognition Letters*, 30(6):623 – 633, April.
- T. Kadir, R. Bowden, E.J. Ong, and A Zisserman. 2004. Minimal training, large lexicon, unconstrained sign language recognition. In *Procs. of BMVC*, volume 2, pages 939 – 948, Kingston, UK, September 7 – 9.
- P. Viola and M Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Procs. of CVPR*, volume 1, pages 511 – 518, Kauai, HI, USA, December.
- Pei Yin, T. Starner, H. Hamilton, I. Essa, and J. M Rehg. 2009. Learning the basic units in american sign language using discriminative segmental feature selection. In *Procs. of ASSP*, pages 4757 – 4760, Taipei, Taiwan, April 19 – 24.

Using ELAN for annotating sign language corpora in a team setting

Onno Crasborn^a & Han Sloetjes^b

^aRadboud University Nijmegen, Centre for Language Studies, PO Box 9103, NL-6500 HD Nijmegen, The Netherlands. E-mail: o.crasborn@let.ru.nl

^bMax Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, the Netherlands. E-mail: han.sloetjes@mpi.nl

Abstract

ELAN is a multimedia annotation tool that is employed in many sign language corpus projects. It is a standalone desktop application that, like many other desktop applications, principally is a single user, document oriented application. In many scenarios this is still perfectly satisfactory but in large-scale corpus projects, involving many collaborators who are working on the same documents, the problem arises of how to resolve edit conflicts and how to prevent undesirable modifications to parts of the document. The Corpus NGT project is such a project and this paper describes the challenges that arose in the process of its creation as well as in the exploitation of this large collection of annotation documents. It outlines recent and possible future development of ELAN and alternate solutions that have been explored and applied.

1. The problem

ELAN is a free, multimodal annotation tool for digital audio and video media. It supports multileveled transcription of up to six synchronized video files per annotation document. The documents are stored as XML (Extensible Markup Language)¹, in its own EAF file format. Over the years, the facilities for working with multiple files have gradually increased. However, in most respects ELAN still assumes that there is a single user for those files, or that users work on the data one at a time. This situation raises several challenges in the creation of large collections of annotation documents that are jointly used by researchers working in a team, as in the case of the development and use of signed language corpora. This paper characterises several of those challenges as they arose in the creation of the *Corpus NGT* and its subsequent exploitation for research. It shows how on the one hand this has steered the recent development of ELAN, and on the other hand complementary solutions have been found that address the complex situation of teamwork on a large set of files. It concludes by suggesting several areas for possible future development of ELAN.

2. Background

2.1 ELAN²

ELAN has a development history of more than 10 years. The software followed the Mac-only application MediaTagger and was called EUDICO in its earliest versions, and it arose from a European project of the latter name.³ The initial set of client-server based viewer applications that were developed in that project, gradually merged into a single standalone annotation editor.

ELAN has originally been, and in fact still is, strongly oriented towards a setting where single users are working on a relatively small number of annotation documents. Like many other desktop applications, and this is probably

true for a majority of them, ELAN assumes that there is only one user at a time working on a document.

At the start of the 21st century, some users expressed their wish to be able to work on annotation documents collaboratively. This led to the implementation of the onsets of a Peer-to-Peer (P2P) based solution for simultaneous, collaborative annotation (Brugman, Crasborn & Russel 2004). In this approach, team members and/or other collaborators are working together at the same time on the same document. Crucial is that the collaborators don't have to be at the same site, sitting at the same workstation. This solution has been implemented and tested up to the demonstration phase, but has never been finalised.

A disadvantage, or at least a limitation, of the above P2P type of collaboration is that the annotators need to be available at the same moment and need to be focussing on the same phenomenon. But in many team situations this is not the most suitable form of collaboration, e.g. in projects where most annotators have specialised into studying a particular kind of phenomena and are working on different tiers in different sections of the media file. One way to handle this, at least in theory, is to let each annotator work in a separate file referring to the same media file(s) and merge all these transcriptions in the end into one complete transcription file using ELAN's "Merge Transcriptions" function. In practice however, this workflow often is not realistic, if only because there is no apparent "end" to the annotation work; it is often not possible to decide when a certain part of the work is finished, and making modifications to a part of the annotation might necessitate re-merging of files. And in some cases it is useful to have the information from annotations on other tiers at hand during the annotation phase (although the opposite can be true as well).

In sections 3 and 4 we describe a combination of solutions that have been created, which consist of a combination of enhancements to ELAN and local solutions for the work with the specific collection that will first be described in the following section.

¹ <http://www.w3.org/standards/xml/>

² <http://www.lat-mpi.eu/tools/elan/>

³ <http://www.mpi.nl/world/tg/lapp/eudico/eudico.html>

2.2 The Corpus NGT⁴

The Corpus NGT is a collection of almost 72 hours of dialogues of 92 different signers for whom NGT is the first language (Crasborn, Zwitserlood & Ros, 2008; Crasborn & Zwitserlood 2008). The recordings for the corpus were created between 2006 and 2008, and the first release of the videos with some initial gloss annotations was published as open content in December 2008. Over 15% of this material received a voice-over from sign language interpreters. A second release of the annotation files including a much larger set of ID-glosses (Johnston 2008) and some sentence-level translations will be published in 2011.

Aside from this publication for linguists as part of the MPI corpus archive, a public version of the data have been made as streaming media in early 2010. The public web site includes a presentation of the data for Deaf people, second language learners, and any interested party. The web site has been translated to German, and an English and NGT version are being planned.

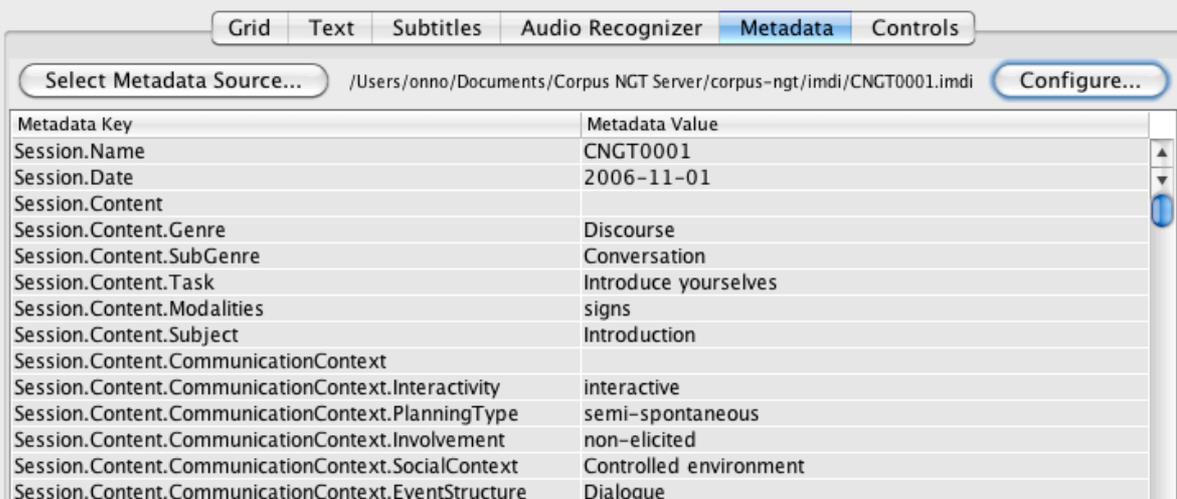
Since its original publication, the 2375 sessions in the Corpus NGT have been used for various research projects. For a project on sign language recognition (SignSpeak), additional gloss annotations are being added and the glosses are being revised to adhere to a more strict ‘one manual form, one gloss’ rule, termed ID-glosses by Johnston (2008). Moreover, for a variety of research projects at Radboud University, many new annotation levels (tiers) have been added. A total of seven researchers and four research assistants regularly add annotations to the corpus now, and perform increasingly complex searches.

3. Working with large sets of annotation documents

The creation of the Corpus NGT involved the segmentation of the data into 2375 parts, each consisting of one annotation file and a number of media files linked to it. Even with a much smaller number of files, it would not be realistic to want to process documents one-by-one: searching or adding tiers only in open documents would not be realistic and would lead to unsystematic files and annotations. For this reason, the Corpus NGT project contributed to the design and implementation of several new functions in ELAN.

The key development in this area was the creation of a link between the metadata descriptions of corpora and the annotation documents. Although ELAN stores some metadata properties of individual annotation documents (such as the ‘Author’ of a document and the ‘Annotator’ of a tier), metadata typically transcend the level of an individual annotation document, classifying sets of documents as sharing the same signers or the same content type or register. Until now, the metadata information that is stored in IMDI files was not accessible from within ELAN. For a search across multiple files with metadata property X, one would have to manually create a domain by selecting annotation documents corresponding to that metadata property one by one in a file selection dialogue, where this information would have to come from another source (such as the IMDI files or another database with the metadata information). Similarly, in order to quickly inspect from which region a participant in the media file comes, one would have to look up the session number in the metadata records.

The first addition that was created to facilitate access to metadata was the creation of a new tab pane in the top right hand part of the ELAN interface. Next to the Grid, Text, Subtitle and Controls pane, a Metadata pane has been created in which the user can select an IMDI file and the fields to be displayed in a table view (Figure 1) or in a



The screenshot shows the ELAN interface with the 'Metadata' tab selected. A 'Select Metadata Source...' button is active, showing the file path: /Users/onno/Documents/Corpus NGT Server/corpus-ngt/imdi/CNGT0001.imdi. A 'Configure...' button is also visible. Below is a table with two columns: 'Metadata Key' and 'Metadata Value'.

Metadata Key	Metadata Value
Session.Name	CNGT0001
Session.Date	2006-11-01
Session.Content	
Session.Content.Genre	Discourse
Session.Content.SubGenre	Conversation
Session.Content.Task	Introduce yourselves
Session.Content.Modalities	signs
Session.Content.Subject	Introduction
Session.Content.CommunicationContext	
Session.Content.CommunicationContext.Interactivity	interactive
Session.Content.CommunicationContext.PlanningType	semi-spontaneous
Session.Content.CommunicationContext.Involvement	non-elicited
Session.Content.CommunicationContext.SocialContext	Controlled environment
Session.Content.CommunicationContext.EventStructure	Dialogue

Figure 1. The Metadata pane in ELAN displays a selection of metadata properties from an IMDI file.

⁴ <http://www.ru.nl/corpusngtuk>

tree view.

Secondly, multiple file searches were enhanced so that they can make use of the output of a search in the IMDI Browser. Thus, in a two-step process one can first search for metadata characteristics and then use the outcome of that search for annotation searches in ELAN. To this end, the IMDI Browser was adapted so that it would be possible to save search results in a file that can then be read by ELAN. The selection of the IMDI file and the specific metadata fields to display is stored in the preferences file.

Most of the available multiple file processes, such as

that allows adding, changing and deleting tiers and linguistic types in multiple documents. Here too, the user can make a selection of a domain to modify and store that domain for later use. The implementation offers a tabular overview of the different tiers and tier properties (Linguistic Type, Annotator, Participant) that are used by all the files in the set, which can help to keep a corpus organised. As users are free to add new tiers to and modify existing ones in any document in a corpus collection, they can also create inconsistencies. These can be easily spotted in the *Multiple File Editor* interface (Figure 2).

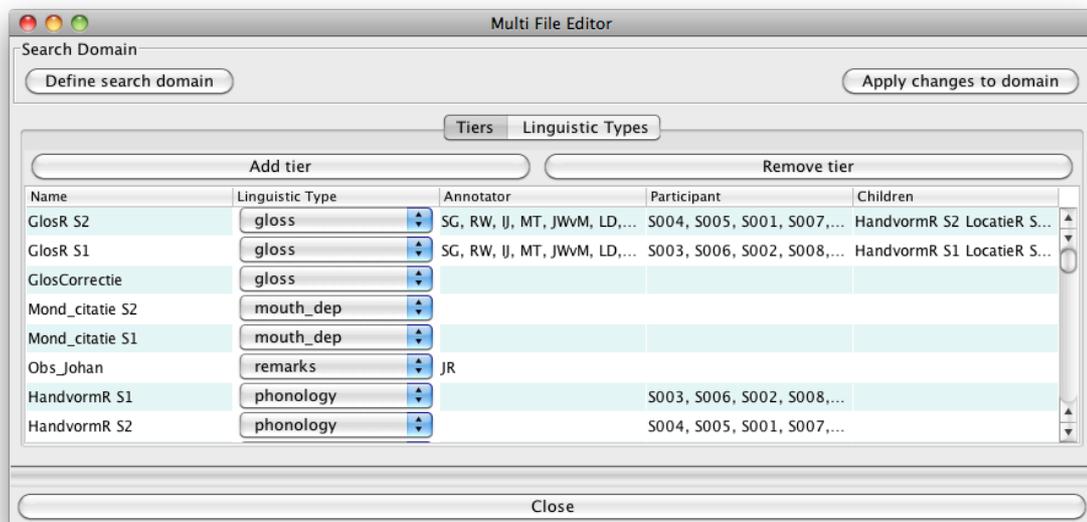


Figure 2. The Multiple Files edit function gives an insightful overview of properties of tiers and their properties.

searching in multiple annotation files, are accompanied by a “domain” selection facility. Domains in this context are selections of files and folders that can be saved in ELAN and reused later. Domains can either be composed manually, by selecting files and folders in a custom file browser window, or they can be derived from an IMDI metadata search result as described above.

Actions that can be performed on multiple files now consist of the following:

- structured search
- find and replace
- generation of statistics
- new document creation based on a template and sets of media files
- annotation “scrubbing” (removal of superfluous spaces, tabs and new lines)
- export as word list, export a selection of tiers and export to tab-delimited text

For all these purposes, then, a selection can be made in the IMDI Browser so that the action would only apply to annotation documents that relate to, for instance, signers from a specific region or from a specific age group.

A special case of processing multiple files is the module

4. Working with a research team on a corpus of annotation files

One of the changes in ELAN made to improve the work in a team setting has been the introduction of the ‘Annotator’ attribute in the specification of tiers. This has been added in ELAN version 3.0; at the same time a corresponding change was made in the EAF schema, in version 2.4. This attribute can be used to sort or group tiers and for creating statistics per annotator. It is expected that the existing “Compare Annotators” function will be extended to make use of this attribute. This function currently produces a rough calculation of the level of agreement between two annotators or raters.

Other tasks were not yet implemented in ELAN at the time of the construction of the Corpus NGT. One function (currently under development) was to create new EAF files based on a template and a list of media files. To facilitate the generation of new documents for the Corpus NGT, Perl scripts were written to create EAF files for a set of media files, and to create PFSX files for a folder of EAF files, based on a dummy PFSX file that was configured to meet specific needs.

In the Corpus NGT annotation documents, specific tiers have been created for exchanging information. There is a *Observations* tier per team member, in which notes for colleagues can be stored. The tier *GlossCorrection* is used for marking possible errors in the glosses, to be double-checked by a team member with that responsibility.

As ELAN is not set up as a client-server system, a solution was sought in which the annotation documents would still be stored in a central space and accessible for all team members. A satisfactory solution until now has been to use the Subversion (SVN) file versioning system, which is typically used in the context of software development in teams. There is a SVN server on the network that creates a backup of every version of every annotation document ever created. When storing a new revision of a file, annotators can add comments as to what was changed in this version of the file. Aside from the backup facility, an advantage of this system is that all users can immediately profit from new annotations as soon as they are uploaded to the server.

The downside of the versioning system is that it imposes heavy demands on the users to stick to strict workflows. Repairing conflicting versions may take quite some time. Moreover, it is not a principled solution: Subversion is really targeted at situations where the text files *themselves* are edited by users, as in software development. In the case of EAF documents, which are an instance of XML, ELAN assumes that there are no other editors of the XML code than ELAN itself, and this can make comparing conflicting versions rather hard. This is particularly so when it comes to the coding of time positions and annotation IDs.

In addition to the EAF files, the SVN server also hosts all the IMDI metadata files and one folder of PFSX files per researcher or research goal. The location of the folder with preferences files can be set in the ELAN preferences since version 3.7.2. Users can thus have access to a uniform ELAN interface for all the documents they open, irrespective of who most recently edited the document.

The applicability of preferences files has been improved by saving preferences when a template file is created. Every new annotation file based on such a template with an associated preferences file, starts with the inherited preferences settings.

5. Areas of further development

The development of a more systematic use of the concept ‘user’ could further facilitate the use of ELAN in teams. Perhaps the possibility of choosing a server-client setup where information about user actions can be systematically stored and conflicts between actions of different users can be prevented would merit consideration again. The iLex tool uses this type of design successfully.⁵ This might entail a shift from an XML document oriented approach to a managed database oriented approach.

⁵ <http://www.sign-lang.uni-hamburg.de/ilex>

There are a number of issues and wishes, brought forward by several user groups, which are seemingly related to the issues discussed in this paper:

- In team settings a need has emerged to “write protect” certain parts of the document for all or most of the annotators.
- Documents that were created based on a template file, can easily become inconsistent when tiers are renamed or deleted.
- Support for a “stand off” treatment of tiers in different transcription files. The tiers of only one of the files should be editable; the other tiers should be read only.

Finding a way to converge these issues and develop, if possible, a single solution is one of the challenges for future developments.

6. Acknowledgements

The software development and the writing of this paper was made possible by grants from the Netherlands Organisation for Scientific Research (NWO, grants 380-70-008 and 276-70-012) and the European Research Council (ERC Starting Researcher Grant 210373 awarded to Onno Crasborn). Developers who contributed to the improvements in ELAN that are described in this paper: Mark Blokpoel (RU), Albert Russel (MPI), Eric Auer (MPI), Alexander Koenig (MPI).

7. References

- Brugman, H., Crasborn, O., Russel, A. (2004) Collaborative Annotation of Sign Language Data with Peer-to-Peer Technology. In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. M.T. Lino et al., eds. Pp. 213-216.
- Crasborn, O. & I. Zwitterlood (2008) The Corpus NGT: an online corpus for professionals and laymen, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd, eds. ELDA, Paris, pp 44-49.
- Crasborn, O., I. Zwitterlood & J. Ros (2008) 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 Nijmegen.
URL: <http://www.ru.nl/corpusngtuk/>
- Johnston, T. (2008) Corpus linguistics and signed languages: no lemmata, no corpus. In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd, eds. ELDA, Paris, pp. 82-87.
- Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, 41(3), 841-849.

SignSpeak – Understanding, Recognition, and Translation of Sign Languages

Philippe Dreuw¹, Jens Forster¹, Yannick Gweth¹, Daniel Stein¹, Hermann Ney¹,
Gregorio Martinez², Jaume Verges Llahi², Onno Crasborn³, Ellen Ormel³, Wei Du⁴,
Thomas Hoyoux⁴, Justus Piater⁴, Jose Miguel Moya⁵, and Mark Wheatley⁶

¹RWTH, Aachen, Germany

dreuw@cs.rwth-aachen.de

⁴ULg, Liege, Belgium

justus.piater@ulg.ac.be

²CRIC, Barcelona, Spain

gregorio.martinez@cric.cat

⁵TID, Granada, Spain

jmm1@tid.es

³RUN, Nijmegen, The Netherlands

o.crasborn@let.ru.nl

⁶EUD, Brussels, Belgium

mark.wheatley@eud.eu

Abstract

The SignSpeak project will be the first step to approach sign language recognition and translation at a scientific level already reached in similar research fields such as automatic speech recognition or statistical machine translation of spoken languages. Deaf communities revolve around sign languages as they are their natural means of communication. Although deaf, hard of hearing and hearing signers can communicate without problems amongst themselves, there is a serious challenge for the deaf community in trying to integrate into educational, social and work environments. The overall goal of SignSpeak is to develop a new vision-based technology for recognizing and translating continuous sign language to text. New knowledge about the nature of sign language structure from the perspective of machine recognition of continuous sign language will allow a subsequent breakthrough in the development of a new vision-based technology for continuous sign language recognition and translation. Existing and new publicly available corpora will be used to evaluate the research progress throughout the whole project.

1. Introduction

The SignSpeak project¹ is one of the first EU funded projects that tackles the problem of automatic recognition and translation of continuous sign language.

The overall goal of the SignSpeak project is to develop a new vision-based technology for recognizing and translating continuous sign language (i.e. provide Video-to-Text technologies), in order to provide new e-Services to the deaf community and to improve their communication with the hearing people.

The current rapid development of sign language research is partly due to advances in technology, including of course the spread of Internet, but especially the advance of computer technology enabling the use of digital video (Crasborn et al., 2007). The main research goals are related to a better scientific understanding and vision-based technological development for continuous sign language recognition and translation:

- understanding sign language requires better linguistic knowledge
- large vocabulary recognition requires more robust feature extraction methods and a modeling of the signs at a sub-word unit level
- statistical machine translation requires large bilingual annotated corpora and a better linguistic knowledge for phrase-based modeling and alignment

Therefore, the SignSpeak project combines innovative scientific theory and vision-based technology development by gathering novel linguistic research and the most advanced techniques in image analysis, automatic speech recognition (ASR) and statistical machine translation (SMT) within a common framework.

1.1. Sign Languages in Europe

Signed languages vary like spoken languages do: they are not mutually understandable, and there is typically one or more signed language in each country.

Although sign languages are used by a significant number of people, only a few member states of the European Union (EU) have recognized their national sign language on a *constitutional* level: Finland (1995), Slovak Republic (1995), Portugal (1997), Czech Republic (1998 & 2008), Austria (2005), and Spain (2007). The European Union of the Deaf (EUD)², a non-research partner in the SignSpeak project, is a European non-profit making organization which aims at establishing and maintaining EU level dialogue with the “hearing world” in consultation and cooperation with its member National Deaf Associations. The EUD is the only organization representing the interests of Deaf Europeans at European Union level. The EUD has 30 full members (27 EU countries plus Norway, Iceland & Switzerland), and 6 affiliated members (Croatia, Serbia, Bosnia and Herzegovina, Macedonia, Turkey & Israel). Their main goals are the recognition of the right to use an indigenous sign language, the empowerment through communication and information, and the equality in education and employment. In 2008, the EUD estimated about 650,000 Sign Language users in Europe, with about 7,000 official sign language interpreters, resulting in approximately 93 sign language users to 1 sign language interpreter (EUD, 2008; Wheatley and Pabsch, 2010). However, the number of sign language users might be much higher, as it is difficult to estimate an exact number – e.g. late-deafened or hard of hearing people who need interpreter services are not always counted as deaf people in these statistics.

¹www.signspeak.eu

²www.eud.eu

1.2. Linguistic Research in Sign Languages

Linguistic research on sign languages started in the 1950s, with initial studies of Tervoort (Tervoort, 1953) and Stokoe (Stokoe et al., 1960). In the USA, the wider recognition of sign languages as an important linguistic research object only started in the 1970s, with Europe following in the 1980s. Only since 1990, sign language research has become a truly world-wide enterprise, resulting in the foundation of the Sign Language Linguistics Society in 2004³. Linguistic research has targeted all areas of linguistics, from phonetics to discourse, from first language acquisition to language disorders.

Vision-based sign language recognition has only been attempted on the basis of small sets of elicited data (Corpora) recorded under lab conditions (only from one to three signers and under controlled colour and brightness ambient conditions), without the use of spontaneous signing. The same restriction holds for much linguistic research on sign languages. Due to the extremely time-consuming work of linguistic annotation, studying sign languages has necessarily been confined to small selections of data. Depending on their research strategy, researchers either choose to record small sets of spontaneous signing which will then be transcribed to be able to address the linguistic question at hand, or native signer intuitions about what forms a correct utterance.

1.3. Research and Challenges in Automatic Sign Language Recognition

In (Ong and Ranganath, 2005; Y. Wu, 1999) reviews on research in sign language and gesture recognition are presented. In the following we briefly discuss the most important topics to build up a large vocabulary sign language recognition system.

1.3.1. Languages and Available Resources

Almost all publicly available resources, which have been recorded under lab conditions for linguistic research purposes, have in common that the vocabulary size, the types/token ratio (TTR), and signer/speaker dependency are closely related to the recording and annotation costs. Data-driven approaches with systems being automatically trained on these corpora do not generalize very well, as the structure of the signed sentences has often been designed in advance (von Agris and Kraiss, 2007), or offer small variations only (Dreuw et al., 2008b; Bungeroth et al., 2008), resulting in probably over-fitted language models. Additionally, most self-recorded corpora consists only of a limited number of signers (Vogler and Metaxas, 2001; Bowden et al., 2004).

For automatic sign language recognition, promising results have been achieved for continuous sign language

recognition under lab conditions (von Agris and Kraiss, 2007; Dreuw et al., 2007a). In the recently very active research area of sign language recognition, a new trend towards broadcast news or weather forecast news can be observed. The problem of aligning an American Sign Language (ASL) sign with an English text subtitle is considered

in (Farhadi and Forsyth, 2006). In (Buehler et al., 2009; Cooper and Bowden, 2009), the goal is to automatically learn a large number of British Sign Language (BSL) signs from TV broadcasts. Due to limited preparation time of the interpreters, the grammatical differences between “real-life” sign language and the sign language used in TV broadcast (being more close to Signed Exact English (SEE)) are often significant. Even if the performances of the automatic learning approaches presented in those works are still quite low, they represent an interesting approach for further research.

1.3.2. Environment Conditions and Feature Extraction

Further difficulties for such sign language recognition frameworks arise due to different environment assumptions. Most of the methods developed assume closed-world scenarios, e.g. simple backgrounds, special hardware like data gloves, limited sets of actions, and a limited number of signers, resulting in different problems in sign language feature extraction or modeling.

1.3.3. Modeling of the Signs

In continuous sign language recognition, as well as in speech recognition, coarticulation effects have to be considered. One of the challenges in the recognition of continuous sign language on large corpora is the definition and modelling of the basic building blocks of sign language. The use of whole-word models for the recognition of sign language with a large vocabulary is unsuitable, as there is usually not enough training material available to robustly train the parameters of the individual word models. A suitable definition of sub-word units for sign language recognition would probably alleviate the burden of insufficient data for model creation.

In ASR, words are modelled as a concatenated sub-word units. These sub-word units are shared among the different word-models and thus the available training material is distributed over all word-models. On the one hand, this leads to better statistical models for the sub-word units, and on the other hand it allows to recognize words which have never been seen in the training procedure using lexica. According to the *linguistic* work on sign language by Stokoe (Stokoe et al., 1960), a phonological model for sign language can be defined, dividing signs into their four constituent visemes, such as the hand shapes, hand orientations, types of hand movements, and body locations at which signs are executed. Additionally, non-manual components like facial expression and body posture are used. However, no suitable decomposition of words into sub-word units is currently known for the purposes of a large vocabulary sign language *recognition* system (e.g. a grapheme-to-phoneme like conversion and use of a pronunciation lexicon).

The most important of these problems are related to the lack of generalization and overfitting systems (von Agris and Kraiss, 2007), poor scaling (Buehler et al., 2009; Cooper and Bowden, 2009), and unsuitable databases for mostly data driven approaches (Dreuw et al., 2008b).

³www.slls.eu

1.4. Research and Challenges in Statistical Machine Translation of Sign Languages

While the first papers on sign language translations only date back to roughly a decade (Veale et al., 1998) and typically employed rule-based systems, several research groups have recently focussed on data-driven approaches. In (Stein et al., 2006), a SMT system has been developed for German and German sign language in the domain weather reports. Their work describes the addition of pre- and post-processing steps to improve the translation for this language pairing. The authors of (Morrissey and Way, 2005) have explored example-based MT approaches for the language pair English and sign language of the Netherlands with further developments being made in the area of Irish sign language. In (Chiu et al., 2007), a system is presented for the language pair Chinese and Taiwanese sign language. The optimizing methodologies are shown to outperform a simple SMT model. In the work of (San-Segundo et al., 2006), some basic research is done on Spanish and Spanish sign language with a focus on a speech-to-gesture architecture.

2. Speech and Sign Language Recognition

Automatic speech recognition (ASR) is the conversion of an acoustic signal (sound) into a sequence of written words.

Due to the high variability of the speech signal, speech recognition – outside lab conditions – is known to be a hard problem. Most decisions in speech recognition are interdependent, as word and phoneme boundaries are not visible in the acoustic signal, and the speaking rate varies. Therefore, decisions cannot be drawn independently but have to be made within a certain context, leading to systems that recognize whole sentences rather than single words.

One of the keys idea in speech recognition is to put all ambiguities into probability distributions (so called stochastic knowledge sources, see Figure 1). Then, by a stochastic modelling of the phoneme and word models, a pronunciation lexicon and a language model, the free parameters of the speech recognition framework are optimized using a large training data set. Finally, all the interdependencies and ambiguities are considered jointly in a search process which tries to find the best textual representation of the captured audio signal. In contrast, rule-based approaches try to solve the problems more or less independently.

In order to design a speech recognition system, four crucial problems have to be solved:

1. preprocessing and feature extraction of the input,
2. specification of models and structures for the words to be recognized,
3. learning of the free model parameters from the training data, and
4. search of the maximum probability over all models during recognition (see Figure 1).

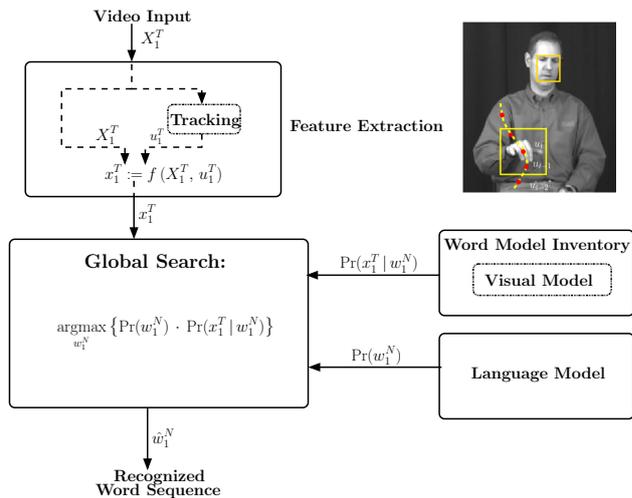


Figure 1: Sign language recognition system overview

2.1. Differences Between Spoken Language and Sign Language

Main differences between spoken language and sign language are due to linguistic characteristics such as simultaneous facial and hand expressions, references in the virtual signing space, and grammatical differences as explained more detailed in (Dreuw et al., 2008c):

Simultaneousness: Major issue in sign language recognition compared to speech recognition – a signer can use different communication channels (facial expression, hand movement, and body posture) in parallel.

Signing Space: Entities like persons or objects can be stored in a 3D body-centered space around the signer, by executing them at a certain location and later just referencing them by pointing to the space – the challenge is to define a model for spatial information handling.

Coarticulation and Epenthesis: In continuous sign language recognition, as well as in speech recognition, coarticulation effects have to be considered. Due to location changes in the 3D signing space, we also have to deal with the movement epenthesis problem (Vogler and Metaxas, 2001; Yang et al., 2007). Movement epenthesis refers to movements which occur regularly in natural sign language in order to move from the end state of one sign to the beginning of the next one. Movement epenthesis conveys no meaning in itself but contributes phonetic information to the perceiver.

Silence: opposed to automatic speech recognition, where the energy of the audio signal is usually used for the silence detection in the sentences, new spatial features and models will have to be defined for silence detection in sign language recognition. Silence cannot be detected by simply analyzing motion in the video, because words can be signed by just holding a particular posture in the signing space over time. Further, the rest position of the hand(s) may be somewhere in the signing space.

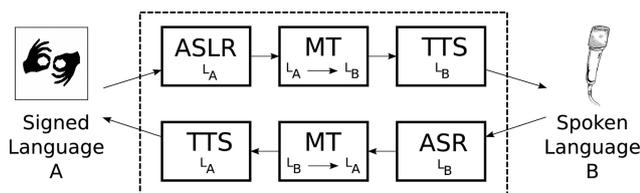


Figure 2: Complete six components-engine necessary to build a Sign-Language-to-Spoken-Language system (components: automatic sign language recognition (ASLR), automatic speech recognition (ASR), machine translation (MT), and text-to-speech/sign (TTS))

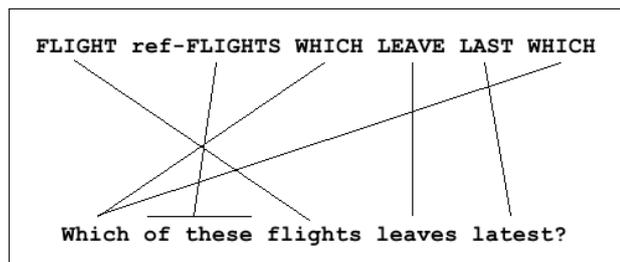


Figure 3: Reorderings are learned by a statistical machine translation system.

3. Towards a Sign-Language-to-Spoken-Language Translation System

The interpersonal communication problem between signer and hearing community could be resolved by building up a new communication bridge integrating components for sign-, speech-, and text-processing. To build a Sign-Language-to-Spoken-Language translator for a new language, a six component-engine must be integrated (see Figure 2), where each component is in principle language independent, but requires language dependent parameters/models. The models are usually automatically trained but require large annotated corpora.

The ASLR recognition is only the first step of a sign-language to spoken-language system. The intermediate representation of the recognized signs is further processed to create a spoken language translation. Statistical machine translation (MT) is a data-based translation method that was initially inspired by the so-called noisy-channel approach: the source language is interpreted as an encryption of the target language, and thus the translation algorithm is typically called a decoder. In practice, statistical machine translation often outperforms rule-based translation significantly on international translation challenges, given a sufficient amount of training data. As proposed in (Stein et al., 2007), a statistical machine translation system is used in SignSpeak to automatically transfer the meaning of a source language sentence into a target language sentence.

As mentioned above, statistical machine translation requires large bilingual annotated corpora. This is extremely important in order to train word reorderings or translate pointing references to the signing space (c.f. Figure 3). As reported in (Dreuw et al., 2007b), novel hand and face tracking features (see Figure 5) will be analyzed and integrated into the SignSpeak translation framework.

In SignSpeak, a theoretical study will be carried out about how the new communication bridge between deaf and hearing people could be built up by analyzing and adapting the ASLR and MT components technologies for sign language processing. The problems described in Section 2. will mainly be tackled by

- analysis of linguistic markers for sub-units and sentence boundaries,
- head and hand tracking of the dominant and non-dominant hand,
- facial expression and body posture analysis,
- analysis of linguistically- and data-driven sub-word units for sign modeling,
- analysis of spatio-temporal across-word modeling,
- signer independent recognition by pronunciation modeling, language model adaptation, and speaker adaptation techniques known from ASR
- contextual and multi-modal translation of sign language by an integration of tracking and recognition features into the translation process

Once the different modules are integrated within a common communication platform, the communication could be handled over 3G phones, media center TVs, or video telephone devices. The following sign language related application scenarios would be possible:

- e-learning of sign language
- automatic transcription of video e-mails, video documents, or video-SMS
- video subtitling

3.1. Impact on Other Industrial Applications

The novel features of such systems provide new ways for solving industrial problems. The technological breakthrough of SignSpeak will clearly have an impact on other applications fields:

Improving human-machine communication by gesture:

vision-based systems are opening new paths and applications for human-machine communication by gesture, e.g. Play Station's EyeToy or Microsoft Xbox's Natal Project⁴, which could be interesting for physically disabled individuals or even blind people as well.

Medical sector: new communication methods by gesture are being investigated to improve the communication between the medical staff, the computer, and other electronic equipments. Another application in this sector is related to web- or video-based *e-Care / e-Health* treatments, or an auto-rehabilitation system which makes the guidance process to a patient during the rehabilitation exercises easier.

⁴www.xbox.com/en-US/live/projectnatal/

Surveillance sector: person detection and recognition of body parts or dangerous objects, and their tracking within video sequences or in the context of quality control and inspection in manufacturing sectors.

4. Available and New Resources Within SignSpeak

All databases presented in this section are either freely available or can be purchased. Depending on the tasks and progress within the SignSpeak project, the focus will be shifted to one of the following databases briefly described in this section. Examples images showing the different recording conditions are shown for each database in Figure 4, where Table 1 gives an overview how the different corpora can be used for evaluation experiments.

4.1. CORPUS-NGT Database

The core of the SignSpeak data will come from the Corpus-NGT⁵ database. This 72 hour corpus of Sign Language of the Netherlands is the first large open access corpus for sign linguistics in the world. It presently contains recordings from 92 different signers, mirroring both the age variation and the dialect variation present in the Dutch Deaf community (Crasborn et al., 2008).

For the SignSpeak project, the limited gloss annotations that were present in the first release of 2008 have been considerably expanded, and sentence-level translations have been added. Furthermore, more than 3000 frames will be annotated to evaluate hand and head tracking algorithms.

4.2. Boston Recordings

All databases presented in this section are freely available for further research in linguistics⁶ and recognition⁷. The data were recorded by Boston University, the database subsets were defined at the RWTH Aachen University in order to build up benchmark databases (Dreuw et al., 2008b) that can be used for the automatic recognition of isolated and continuous sign language, respectively.

The RWTH-BOSTON-50 database was created for the task of isolated sign language recognition (Zahedi et al., 2006). It has been used for nearest-neighbor leaving-one-out evaluation of isolated sign language words.

The RWTH-BOSTON-104 has been used successfully for continuous sign language recognition experiments (Dreuw et al., 2007a). For the evaluation of hand tracking methods in sign language recognition systems, the database has been annotated with the signers' hand and head positions. More than 15.000 frames in total are annotated and are freely available⁸.

For the task of sign language recognition and translation, promising results on the publicly available benchmark database RWTH-BOSTON-104 have been achieved for automatic sign language recognition (Dreuw et al., 2007a)

and translation (Dreuw et al., 2008c; Dreuw et al., 2007b) that can be used as baseline reference for other researchers. However, the preliminary results on the larger RWTH-BOSTON-400 database show the limitations of the proposed framework and the need for better visual features, models, and corpora (Dreuw et al., 2008b).

4.3. Phoenix Weather Forecast Recordings

The RWTH-PHOENIX database with German sign language annotations of weather-forecast news has been first presented in (Stein et al., 2006) for the purpose of sign language translation (referred to as RWTH-PHOENIX-v1.0 in this work). It consists of about 2000 sentences, 9.000 running words, with a vocabulary size of about 1700 signs. Although the database is suitable for recognition experiments, the environment conditions in the first version cause problems in robust feature extraction such as hand tracking (see also Figure 4). During the SignSpeak project, a new release RWTH-PHOENIX-v2.0 will be recorded and annotated to meet the demands described in Section 5.. Due to the easier environment conditions in the RWTH-PHOENIX-v2.0 version (see also Figure 4), promising feature extraction and recognition results are expected.

4.4. The ATIS Sign Language Corpus

The ATIS Irish sign language database (ATIS-ISL) has been presented in (Bungeroth et al., 2008), and is suitable for recognition and translation experiments. The Irish sign language corpus formed the first translation into sign language of the original ATIS data. The sentences from the original ATIS corpus are given in written English as a transcription of the spoken sentences. The database as used in (Stein et al., 2007) contains 680 sentences with continuous sign language, has a vocabulary size of about 400 signs, and contains several speakers. For the SignSpeak project, about 600 frames have been annotated with hand and head positions to be used in tracking evaluations.

4.5. SIGNUM Database

The SIGNUM database⁹ has been first presented in (von Agris and Kraiss, 2007) and contains both isolated and continuous utterances of various signers. This German sign language database is suitable for signer independent continuous sign language recognition tasks. It consists of about 33k sentences, 700 signs, and 25 speakers, which results in approximately 55 hours of video material.

5. Experimental Results and Requirements

In order to build a Sign-Language-to-Spoken-Language translator, reasonably sized corpora have to be created for the data-driven approaches. For a limited domain speech recognition task (Verbmobil II) as e.g. presented in (Kanthak et al., 2000), systems with a vocabulary size of up to 10k words have to be trained with at least 700k words to obtain a reasonable performance, i.e. about 70 observations per vocabulary entry. Similar values must be obtained for a limited domain translation task (IWSLT) as e.g. presented in (Mauser et al., 2006).

⁵www.corpusngt.nl

⁶<http://www.bu.edu/asllrp/>

⁷<http://www-i6.informatik.rwth-aachen.de/aslr/>

⁸www-i6.informatik.rwth-aachen.de/~dreuw/database.php

⁹<http://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/>



Figure 4: Example images from different corpora used in SignSpeak (f.l.t.r.): Corpus-NGT, RWTH-BOSTON, RWTH-PHOENIX v1.0 and v2.0, ATIS-ISL, and SIGNUM

Table 1: Sign language corpora used within SignSpeak and their application areas

Corpus	Evaluation Area			
	Isolated Recognition	Continuous Recognition	Tracking	Translation
Corpus-NGT	✓	✓	✓	✓
RWTH-BOSTON-50	✓	✗	✗	✗
RWTH-BOSTON-104	✗	✓	✓	✗
RWTH-BOSTON-400	✗	✓	✗	✗
RWTH-PHOENIX-v1.0	✓	✓	✗	✓
RWTH-PHOENIX-v2.0	✗	✓	✗	✓
ATIS-ISL	✗	✓	✓	✓
SIGNUM	✓	✓	✗	✗

Similar corpora statistics can be observed for other ASR or MT tasks. The requirements for a sign language corpus suitable for recognition and translation can therefore be summarized as follows:

- annotations should be domain specific (i.e. broadcast news, or weather forecasts, etc.)
- for a vocabulary size smaller than 4k words, each word should be observed at least 20 times
- the singleton ratio should ideally stay below 40%

Existing corpora should be extended to achieve a good performance w.r.t. recognition and translation (Forster et al., 2010). During the SignSpeak project, the existing RWTH-PHOENIX corpus (Stein et al., 2006) and Corpus-NGT (Crasborn et al., 2008) will be extended to meet these demands (see Table 2). Novel facial features (Piater et al., 2010) developed within the SignSpeak project are shown in Figure 5 and will be analyzed for continuous sign language recognition.

6. Acknowledgements

This work received funding from the European Community’s Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).

7. References

- R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady. 2004. A Linguistic Feature Vector for the Visual Interpretation of Sign Language. In *ECCV*, volume 1, pages 390–401, May.
- Patrick Buehler, Mark Everingham, and Andrew Zisserman. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *IEEE CVPR*, Miami, FL, USA, June.
- Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The ATIS Sign Language Corpus. In *LREC*, Marrakech, Morocco, May.
- Y.-H. Chiu, C.-H. Wu, H.-Y. Su, and C.-J. Cheng. 2007. Joint Optimization of Word Alignment and Epenthesis Generation for Chinese to Taiwanese Sign Synthesis. *IEEE Trans. PAMI*, **29**(1):28–39.
- Helen Cooper and Richard Bowden. 2009. Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition. In *IEEE CVPR*, Miami, FL, USA, June.
- Onno Crasborn, Johanna Mesch, Dafydd Waters, Annika Nonhebel, Els van der Kooij, Bencie Woll, and Brita Bergman. 2007. Sharing sign language data online. Experiences from the ECHO project. *International Journal of Corpus Linguistics*, **12**(4):537–564.
- 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. Technical report, Centre for Language Studies, Radboud University Nijmegen. <http://www.corpusngt.nl>.
- P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007a. Speech Recognition Techniques for a Sign Language Recognition System. In *ICSLP*, Antwerp, Belgium, August. Best paper award.
- P. Dreuw, D. Stein, and H. Ney. 2007b. Enhancing a Sign Language Translation System with Vision-Based Features. In *Intl. Workshop on Gesture in HCI and Simulation 2007*, pages 18–19, Lisbon, Portugal, May.
- Philippe Dreuw, Jens Forster, Thomas Deselaers, and Hermann Ney. 2008a. Efficient Approximations to Model-based Joint Tracking and Recognition of Continuous Sign Language. In *IEEE International Conference Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, September.
- Philippe Dreuw, Carol Neidle, Vassilis Athitsos, Stan

Table 2: Expected corpus annotation progress of the RWTH-PHOENIX and Corpus-NGT corpora in comparison to the limited domain speech (Vermobil II) and translation (IWSLT) corpora.

	BOSTON-104		Phoenix		Corpus-NGT		Vermobil II	IWSLT	
year	2007	2009	2011	2009	2011		2000	2006	
recordings	201	78	400	116	300		-	-	
running words	0.8k	10k	50k	30k	80k		700k	200k	
vocabulary size	0.1k	0.6k	2.5k	3k	> 5k		10k	10k	
T/T ratio	8	15	20	10	< 20		70	20	
Performance	11% WER (Dreuw et al., 2008a)		-	-	-	-	15% WER (Kanthak et al., 2000)		40% TER (Mauser et al., 2006)

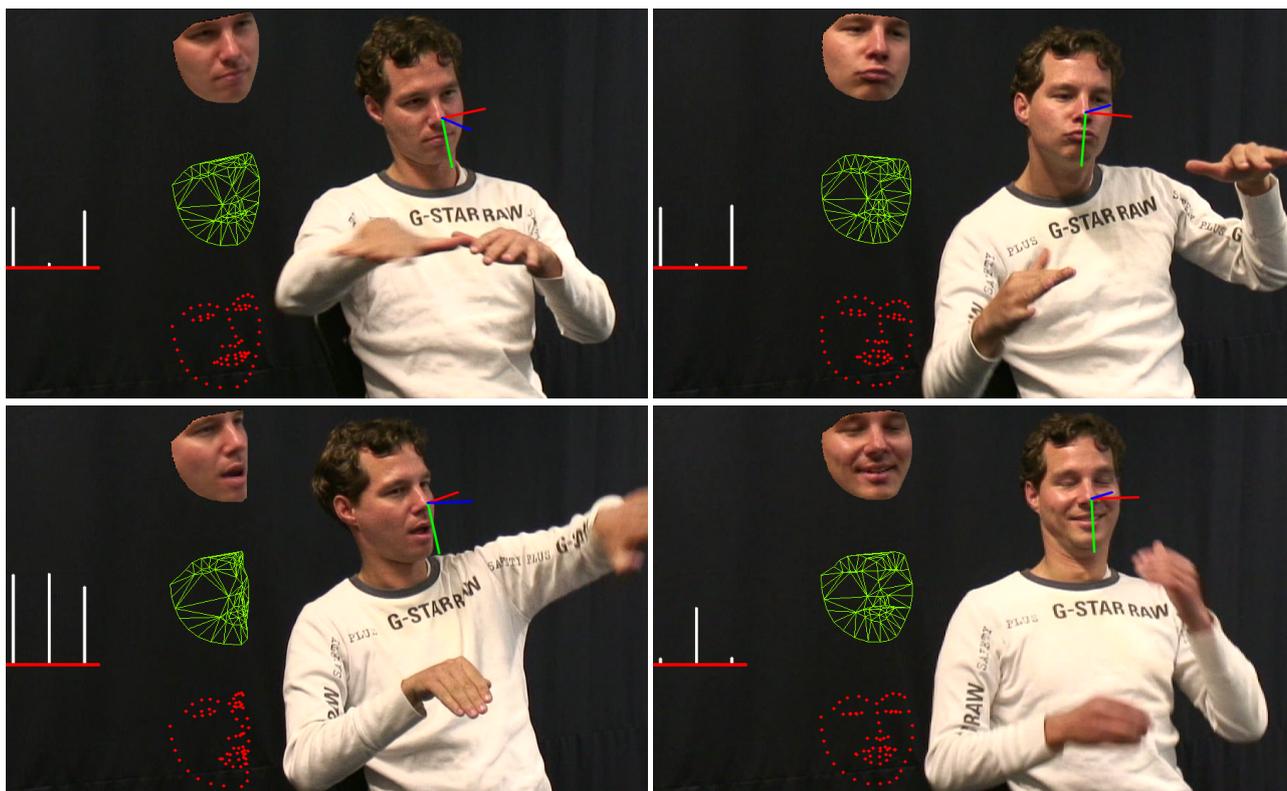


Figure 5: Novel facial feature extraction on the Corpus-NGT database (f.l.t.r.): three vertical lines quantify features like left eye aperture, mouth aperture, and right eye aperture; the extraction of these features is based on a fitted face model, where the orientation of this model is shown by three axis on the face: red is X, green is Y, blue is Z, origin is the nose tip.

- Sciaroff, and Hermann Ney. 2008b. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In *LREC*, Marrakech, Morocco, May.
- Philippe Dreuw, Daniel Stein, Thomas Deselaers, David Rybach, Morteza Zahedi, Jan Bungeroth, and Hermann Ney. 2008c. Spoken Language Processing Techniques for Sign Language Recognition and Translation. *Technology and Disability*, 20(2):121–133, June.
- EUD. 2008. Survey about Sign Languages in Europe.
- A. Farhadi and D. Forsyth. 2006. Aligning ASL for statistical translation using a discriminative word model. In *IEEE CVPR*, New York, USA, June.
- Jens Forster, Daniel Stein, Ellen Ormel, Onno Crasborn, and Hermann Ney. 2010. Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation. In *4th LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Malta, May.
- Stephan Kanthak, Achim Sixtus, Sirko Molau, Ralf Schlüter, and Hermann Ney, 2000. *Fast Search for Large Vocabulary Speech Recognition*, chapter "From Speech Input to Augmented Word Lattices", pages 63–78. Springer Verlag, Berlin, Heidelberg, New York, July.
- Arne Mauser, Richard Zens, Evgeny Matusov, Saša Hasan, and Hermann Ney. 2006. The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation. In *IWSLT*, pages 103–110, Kyoto, Japan, November. Best Paper Award.
- S. Morrissey and A. Way. 2005. An Example-based Approach to Translating Sign Language. In *Workshop in Example-Based Machine Translation (MT Summit X)*, pages 109–116, Phuket, Thailand, September.
- S. Ong and S. Ranganath. 2005. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. *IEEE Trans. PAMI*, 27(6):873–891, June.
- Justus Piater, Thomas Hoyoux, and Wei Du. 2010. Video Analysis for Continuous Sign Language Recognition. In

- 4th LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Malta, May.
- R. San-Segundo, R. Barra, L. F. D'Haro, J. M. Montero, R. Córdoba, and J. Ferreiros. 2006. A Spanish Speech to Sign Language Translation System for assisting deaf-mute people. In *ICSLP*, Pittsburgh, PA, September.
- D. Stein, J. Bungeroth, and H. Ney. 2006. Morpho-Syntax Based Statistical Methods for Sign Language Translation. In *11th EAMT*, pages 169–177, Oslo, Norway, June.
- D. Stein, P. Dreuw, H. Ney, S. Morrissey, and A. Way. 2007. Hand in Hand: Automatic Sign Language to Speech Translation. In *The 11th Conference on Theoretical and Methodological Issues in Machine Translation*, Skoevde, Sweden, September.
- W. Stokoe, D. Casterline, and C. Croneberg. 1960. *Sign language structure. An outline of the visual communication systems of the American Deaf (1993 Reprint ed.)*. Silver Spring MD: Linstok Press.
- B. Tervoort. 1953. Structurele analyse van visueel taalgebruik binnen een groep dove kinderen.
- T. Veale, A. Conway, and B. Collins. 1998. The Challenges of Cross-Modal Translation: English to Sign Language Translation in the ZARDOZ System. *Journal of Machine Translation*, 13, No. 1:81–106.
- C. Vogler and D. Metaxas. 2001. A Framework for Recognizing the Simultaneous Aspects of American Sign Language. *CVIU*, 81(3):358–384, March.
- U. von Agris and K.-F. Kraiss. 2007. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. In *Gesture in Human-Computer Interaction and Simulation*, Lisbon, Portugal, May.
- Mark Wheatley and Annika Pabsch. 2010. Sign Language in Europe. In *4th LREC Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT)*, Malta, May.
- T.S. Huang Y. Wu. 1999. Vision-based gesture recognition: a review. In *Gesture Workshop*, volume 1739 of *LNCS*, pages 103–115, Gif-sur-Yvette, France, March.
- Ruiduo Yang, Sudeep Sarkar, and Barbara Loeding. 2007. Enhanced Level Building Algorithm to the Movement Epenthesis Problem in Sign Language. In *CVPR*, MN, USA, June.
- Morteza Zahedi, Philippe Dreuw, David Rybach, Thomas Deselaers, Jan Bungeroth, and Hermann Ney. 2006. Continuous Sign Language Recognition - Approaches from Speech Recognition and Available Data Resources. In *LREC Workshop on the Representation and Processing of Sign Languages: Lexicographic Matters and Didactic Scenarios*, pages 21–24, Genoa, Italy, May.

Corpus Design for Signing Avatars

Kyle Duarte, Sylvie Gibet

Université de Bretagne-Sud, Laboratoire VALORIA, Vannes, France
 kyle.duarte@univ-ubs.fr, sylvie.gibet@univ-ubs.fr

Abstract

The SignCom project uses motion capture (mocap) data to animate a virtual French Sign Language (LSF) signer. An important part of any signing avatar project is to ensure that a computer animation engine has a large quantity of interesting and on-topic signs from which to build novel signing sequences. In this article, we detail the process of selecting an adequate range of signs and situations to be included in our corpus: from controlling discourse topic to including signs that can accept modified movements or handshapes, we describe how an avatar corpus has a different motivation than traditional signed language corpora.

1. Introduction

Though the field of signed language corpus building is young, designing a corpus specifically for application in signing avatars already requires a deviation from available standard practices. Often, corpora for sign retrieval are based on semi-scripted interactions that yield many instances of a restricted set of signs, useful for building unique dialogs later on. We describe here the considerations we have taken in designing the SignCom signing avatar corpus and how they might vary from corpora designed solely for linguistic analysis.

2. Previous Research

Sinclair defines a computer corpus as “a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks.” These structures are evident in the Australian, British, Dutch, Greek, and other signed language corpora: language samples are coded (usually with ELAN) to indicate phonology, morphosyntax, and other language components for later retrieval and analysis (Johnston and de Beuzeville, 2009; Crasborn and Zwitserlood, 2008; Efthimiou and Fotinea, 2007).

However, where these corpora have been developed to serve as living representations of signed languages across individuals and time, corpora used in the realm of language synthesis attempt to find a restricted sample of language that can be reused in a variety of scenarios.

Akin to digital motion databases that might, for example, index the movements of a basketball player, signing avatar corpora require many repetitions of the same sign in different contexts to provide an interesting base for research and animation. With multiple phonological instances of the same sign recorded, a computer animator can choose a best-fit sign out of many, instead of forcing a single instance of the sign into a novel context. These principles have shaped the range of French Sign Language (LSF) signs made available in the SignCom project.

3. SignCom Corpus Design

The SignCom corpus has been designed by a team of researchers that includes linguists and computer scientists, hearing and Deaf. With multiple points of view converging on solving a multidisciplinary problem, several opposing goals have had to be weighed for our desired outcome.

Three excerpts from the segments we have used most often for language synthesis to date are shown below; after, follow descriptions of our opposing goals and our eventual solutions.

Last Saturday evening I organized a cocktail party. I invited some friends over to my house. In order to facilitate communication, I pushed the chairs in the living room into a semi-circle. There was a coffee table for our drinks, and an American bar with various drinks, fruits, glasses, and straws.

I asked my friend, “what do you want?”
 (S)he said, “I would like vodka and orange juice.”
 “Okay,” I responded. I selected a tall thin glass and added vodka about a quarter of the way up. I filled the rest of the glass with orange juice and handed it to my friend.

I asked the next friend what (s)he wanted.
 (S)he responded, “eh, I like any drink, so I don’t really know. What do you suggest?”
 “I’d suggest a cocktail named *Cuba Libre*,” I said.
 “What’s inside that?” (s)he asked.

I said it would be a surprise. I got a tall glass and added a couple of ice cubes. I poured a little lemon juice in the glass, added some rum to that, and filled the glass with cola, then served it to my friend.

All in all, I was quite happy that the evening went well.

3.1. Depth vs. Breadth and Variation vs. Consistency

Traditional corpora attempt to gather a large number of signs to represent the largest slice possible of a language. For the purposes of language synthesis, however, the researcher wants to have control over the types that appear in the corpus, and would prefer several tokens of these types. Dialogues are thus preplanned to ensure multiple instances of a single type are available for searching and retrieval in later experiments, also allowing for best match selection among token candidates.

The SignCom corpus contains three thematic sections: the Cocktail story, and the Galette and Salad interactions.

These themes limit the material that can be discussed in elicitation sessions to a narrow vocabulary. Discussed in long interactions, the signer provides a large number of tokens relative to the narrow focus. The Cocktail story section, measuring roughly one third of the overall corpus, contains the tokens shown in Table 1, among others.

With this variety and frequency of cocktail-related lexemes, we are able to produce a number of novel utterances around the same subject. For example, Figure 1 shows a sequence we have constructed from various single signs and sign phrases. The final result is interpreted as

I asked the next friend what (s)he wanted.
 (S)he responded, “eh, I don’t like fruity drinks, so I don’t really know. What do you suggest?”
 “I’d suggest a cocktail named *Cuba Libre*,” I said.
 I gave it to her and (s)he took it.
 “Great!”

Note that constructing this utterance requires selecting signs from various parts of the corpus. The movements of two signs were inverted phonologically to evoke a contrary meaning. The purposeful inclusion of such directional signs was intended for such an utterance, and is detailed in Section 3.3., below.

Finally, as there is a necessary balance of control within variability for avatar projects, signing avatar corpora do not provide the level of variation needed for a sociolinguistic study.

Table 1: The tokens of highest occurrence in the Cocktail story section of the SignCom corpus.

14x WHAT	7x WANT
9x VARIOUS	4x FILL
8x COCKTAIL	3x JUICE
8x DRINK (n.)	3x ORANGE
8x EVENING	3x VODKA
8x FRUIT	2x RUM
8x POUR	2x SUGGEST
7x GLASS	

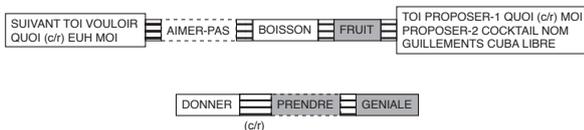


Figure 1: Signs can be rearranged to create novel phrases. Here, signs are retrieved from two different recording takes (white and gray backgrounds) and linked with transitions created by our animation engine (striped background). The sign AIMER (“like”) is reversed to create AIMER-PAS (“dislike”), as is DONNER (“give”) to create PRENDRE (“take”). Finally, a role shift, shown as (c/r), is included in one transition to ensure discourse accuracy and comprehension.

3.2. Open-Ended vs. Scripted

Anonymity in contributions to signed language corpora has been an important conversation within the Deaf communities that support this type of research. At the most basic level, given the face’s active involvement in the signing event it is impossible to hide the identity of the signer. Linguistic data has thus been subject to tight controls regarding rights releases to allow data analysis among researchers, as well as data publishing to wider and/or public audiences.

This topic becomes even more sensitive when open-ended questions are used to elicit stories for linguistic corpora. Existing corpora use guiding topics to elicit personal responses, which may include reports of abuse or other illegal activities; eventually such data would require censorship when making corpora public. As signing avatars almost inevitably become publicly viewable, researchers aim to avoid controversial topics in recording sessions.

As an added benefit, the avatar medium aides in anonymizing elicited data by providing a new face and body for the signer. Figure 2 shows our language consultant alongside the avatar that replays her signing in our animation system.



Figure 2: Avatars provide new identities to signers without covering the face, an important articulator for the signing event.

3.3. Experiments in altering phonological components

Our specific research interests brought us to include a number of indicating verbs and depicting verbs in the SignCom corpus. Among our scientific inquiries are the questions of whether playing reversible indicating verb motions backwards will be convincing and whether altering the handshape of a stored depicting verb will be understood as a change in meaning.

For example, an LSF signer can reverse the movement of the LSF sign AIMER (“like”) to produce the meaning

AIMER-PAS (“dislike”), however human motion theories predict that the motion profile of the reversed sign AIMER-PAS will not be a simple inversion of the forward sign AIMER. By creating sequences that include the captured sign AIMER played backward in a computer-generated utterance, we can perform simple perception tests with signers to judge the acceptability of this relatively straightforward animation technique. We believe these inversions will be understood by signers, so we have included them in our corpus to challenge existing understandings of human motion.

In addition, given our inclusion of signs that take multiple handshapes, like DONNER (“give”), we can substitute handshapes from other signs to influence meaning. In the case of DONNER, most often sign in our corpus as if the signer is handing a glass to someone, a handshape substitution could yield additional meanings, such as giving a piece of paper or giving something dirty (paired with an appropriate facial expression).

3.4. Technical Considerations

Finally, avatar systems must incorporate motion capture (mocap) files that represent the movement of the body, generally much more compact than video files. This incorporation, as well as results of our avatar corpus, is detailed in the paper “Heterogeneous Data Sources for Signed Language Analysis and Synthesis” presented at the LREC 2010 main conference (Duarte and Gibet, 2010).

4. Conclusion

In all, creating databases of signs for signing avatars is not unlike some aspects of traditional linguistic corpora. However, key factors such as dialogue content and style, as well as technical inclusions, must be considered in designing an avatar corpus.

For the SignCom project, we have centered our elicitation sessions around three themes so as to limit the scope of vocabulary attained, and increase the tokens available to us for creating similarly-themed novel utterances. By studying semi-scripted stories, we virtually eliminate the possibility that the signer provides sensitive information that should be held from the public’s view, and better control the corpus’s content for later retrieval. By the nature of animating an avatar, we preserve anonymity for our signer.

Other project goals brought us to include a number of signs that could exist with altered movements or handshapes, to test our animation system’s ability to interchange body parts across signs, as well as to better understand signers’ perception and comprehension of signing avatars.

Having collected our data, we believe that we have an excellent base with which we can create convincing animations of French Sign Language, due in large part to the intentional way we built the SignCom corpus.

5. Acknowledgements

We would like to thank Julia Pelhate (Websourd, Toulouse) and Juliette Dalle for their participation in our signed language mocap recordings, as well as Patrice Dalle (IRIT, Toulouse) for his support on the corpus dialogue design.

This project is funded by the *French National Research Agency*.

6. References

- Onno Crasborn and Inge Zwitterlood. 2008. Annotation of the video data in the Corpus NGT. Technical report, Department of Linguistics and Center for Language Studies, Radboud University, Nijmegen, the Netherlands, November.
- Kyle Duarte and Sylvie Gibet. 2010. Heterogeneous data sources for signed language analysis and synthesis. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta*, 19-21 May. To appear.
- Eleni Efthimiou and Stavroula-Evita Fotinea. 2007. GSLC: Creation and annotation of a Greek Sign Language corpus for HCI. In *Universal Access in Human Computer Interaction. Coping with Diversity*, volume 4554 of *Lecture Notes in Computer Science*, pages 657–666. Springer, Berlin.
- Trevor Johnston and Louise de Beuzeville. 2009. Researching the linguistic use of space in Auslan: Guidelines for annotators using the Auslan corpus. Technical report, Department of Linguistics, Macquarie University, Sydney, June.

Towards decoding Classifier function in GSL

Eleni Efthimiou, Stavroula-Evita Fotinea, Athanasia-Lida Dimou, and Constandinos Kalimeris

Institute for Language and Speech Processing
Artemidos 6 & Epidavrou, 151 25 Maroussi, Athens, Greece
{eleni_e, evita, ndimou, c_kal}@ilsp.gr

Abstract

Here we will present work based on a corpus specially designed and elicited in order to provide data for the study of classifier function in Greek Sign Language (GSL). Data elicitation was based on presentation to informants of a series of stimuli which lead to utterances entailing the set of classifier functions met in the language. The whole set of video recorded data were annotated in order to provide an appropriate corpus for the investigation of classifier instantiations. Annotation work was complemented by the use of a search tool, external to the ELAN environment, that allows to create a data base of annotated video clips by exploiting the set of classification features used to annotate the video recorded data. Theoretical analysis of the so created linguistic data supported the formulation of a proposal for classifier behaviour which differentiates among three distinguished grammar functions based on the property of classifiers to act as semantic markers that create semantic classes of objects sharing common semantic features.

1. Introduction

Video storage of linguistic data has allowed for the application of corpus based approaches to linguistic analysis, which are only recently been made possible.

In this paper we propose an analysis of GSL classifiers focusing on the realisation of Classifier Predicates (CP) as distinct pronoun morphemes, albeit attached as clitics to the base morpheme denoting the predicator (the “verb”) of the CP. The use of classifiers is predominant in GSL, similar to other known SL systems. In the current study, we focused on identifying all instantiations of classifier function in the GSL system, in order to support a theoretical account covering the spectrum of classifier uses, spanning from their appearance as bound morphemes of semantic class on base signs, up to bounding elements in co-indexing. To serve the theoretical study, a special corpus has been elicited and properly annotated. The current study was triggered by the lack of a systematic definition of classifier use in GSL, and became necessary in the framework of a grammar model for the theoretical analysis of the language.

2. Classifier corpus elicitation & annotation

2.1 Corpus elicitation

In order to collect appropriate data for the reported study, a purpose-driven set of visual stimuli to be presented to natural signers was created (figure 1). The stimuli were divided to three categories. The first category was composed of pictures of a) human beings executing specific actions or having specific body postures, and b) arrangements of objects of varying shapes and sizes, either grouped according to shape similarity or following spatial arrangements of geometrical nature. The second category of stimuli entailed the task of narration of different stories on the basis of sets of pictures triggering the use of classifiers during signing of depicted action. The third category involved cartoon animation, which

after been watched, the signers were asked to provide a detailed summary of the displayed action. Each informant was presented with the same complete set of visual stimuli and was video recorded while signing the related tasks. The so elicited data provided a corpus which contains significant instantiations of classifier use in GSL. In order to exploit the material of the corpus, an annotation procedure was applied, as described next.

2.2 Corpus annotation

The content of the video corpus produced through the above mentioned elicitation method was annotated according to the following four annotation tiers (figure 2):

- a) “Discourse Unit”: in this tier we annotated the content of the video, clustered into ample categories, which correspond to the visual stimuli provided during the elicitation procedure, i.e. “various types of tables”, “various types of cups” etc.
- b) “CP_MAX”: in this tier we have marked the maximal CP signed by the informant. This is a subunit of the “Discourse Unit” tier and refers to the immediate semantic content of classifiers used in signing utterances, i.e. “round tables of different size”, “pipes of different dimension” etc.
- c) “CP_GLOSS”: this is the tier mostly exploited in our study at the current stage of research work. Each sign phrase annotated with a “CP_MAX” value is split into its respective constituents; the latter being values for “CP_GLOSS”, which may correspond to either signs or classifiers, including annotation strings such as “table”, “round”, “SIZE” etc to indicate the related semantic content.
- d) “HS”: in this tier font symbols indicate the handshape or handshapes involved in the signing of each “CP_GLOSS”, i.e. “D”, “L”, “b” etc.

These four tiers provide the necessary information to group pieces of data as to the different classifiers and classifier functions met in GSL. Our interest focuses on the ability of classifier morphemes to a) create new lexicon items when combined with individual signs, b) add qualitative/quantitative values to entities, and c) serve

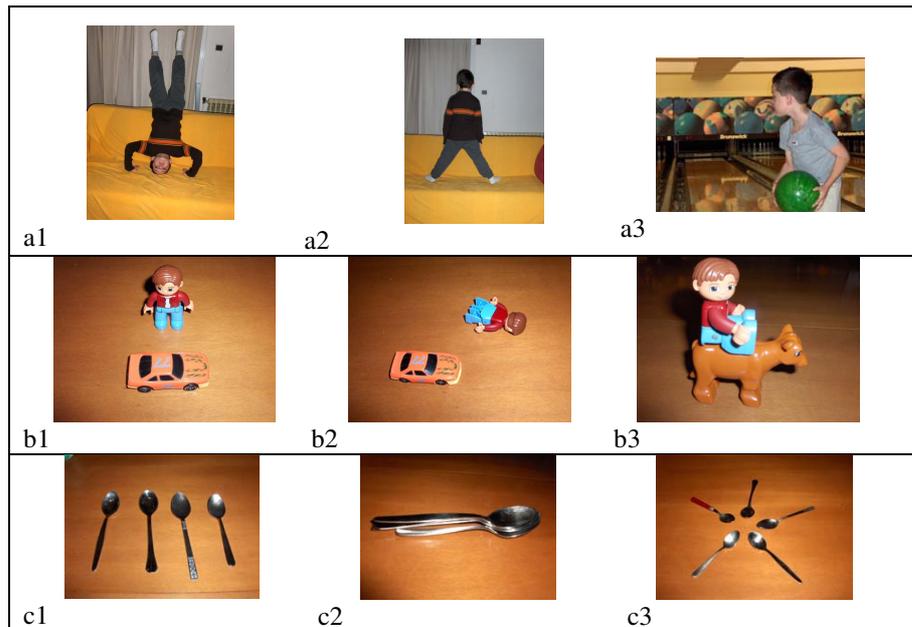


Figure 1: Sample of visual stimuli for the elicitation of the corpus

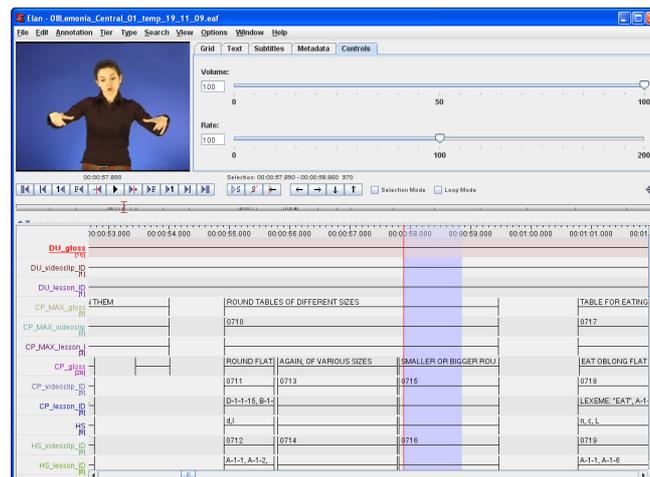


Figure 2: Corpus annotation

co-indexing within phrase utterance.

In order to apply annotation markings which would reveal classifier functions, prior to annotation work, a coding scheme based on four major categories of attributes was adopted. The annotation categories -coded as A, B, C, D followed by 2 up to 4 digits to indicate specific subcategories- were used for the annotation of the “CP_GLOSS” tier, when the latter involved a classifier rather than an independent lexical item. The four annotation categories are sketched below:

- A: a rough ontological division was made into human and non-human entities. In this respect, the coded categories A-1 correspond to different kinds of objects, their description relying merely to their shape, while the A-2 categories refer to humans and the respective subcategories to parts of the human body.
- B: it describes the relevant position of an entity. Subcategories B-1 describe static relevant positions (in front of something or someone (sth/smn), on top of sth/smn, etc), while subcategories B-2 refer to

positions that describe the simultaneous presence of another entity (i.e. lining up behind others, following sth/smn, etc). Subcategories B-2 are used in annotation in those case where the signer makes use of both hands; a condition that is not prerequisite for the B-1 case.

- C: it describes the relevant movement of an entity (i.e. downwards, upwards, back and forth, etc.).
- D: it entails descriptions of size relative to shape. This category directly relates to category A, as the iconicity properties of the signed entity which incorporates a classifier, are those dictating the way “size” has to be signed in each case.

In the early stage of the research, the total number of quantised subcategories to be used in annotation reached up to 60. In order to fully define each classifier instantiation, several of annotation subcategories were attributed to one classifier entry. This unavoidable option for annotation has proven less efficient as annotation process progressed since it became hard to manage the

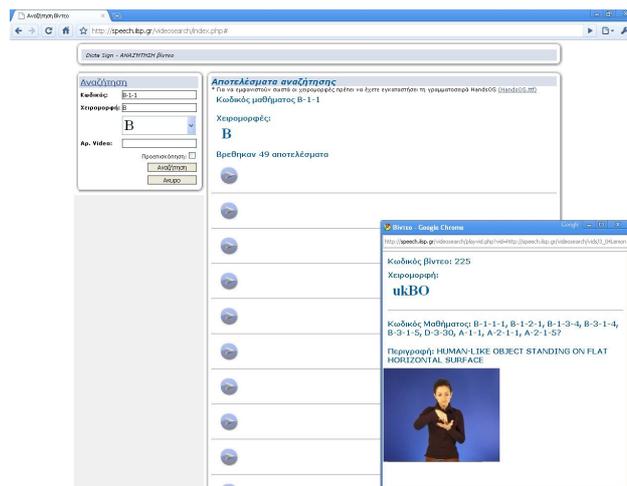


Figure 3: Video Search Tool. Result of a combined search for two annotation codes with retrieved video clips listed. Presentation of a selected item on superimposed window on the right hand side

coded content of the annotated linguistic data.

In order to exploit the patterns of overlapping categories and eventually eliminate redundant ones, we complemented our annotation work with the use of a search tool external to the ELAN environment.

3. Annotated corpus search tool

The annotations search tool is a web based application accessible by <http://speech.ilsp.gr/videosearch/index.php>.

The tool allows extraction of video parts annotated for “CP_GLOSS” values and their storage as individual videos clips. The tool provides for three search options:

- Code: with this search option the user executes simple or combined searches for videos containing one or more annotation codes (i.e. A-1-1, B-2-2). The search result is a list of the videos annotated for the searched code(s) (figure 3).
- Handshape: the search tool facilitates combined search of annotation codes and the handshapes used in classifier formation. This is particularly helpful as the information of the handshape of a Classifier can disambiguate seemingly similar videos and indicate errors during the annotation procedure.
- Video Clip ID: each video clip has a unique identifier number; this search field allows the user to retrieve individual clips that may have caught his/her attention and compare them to one another.

The search tool has proven to be a valuable asset for the present study as it facilitated identification of the characteristics of classifiers, which led to a considerable narrowing down of the initial 60 annotation subcategories, also accelerating the annotation process.

4. Grammatical functions of classifiers

Studies of the syntactic structure of SL utterances reveal systematic patterns. Our corpus-based study of the Greek Sign Language (GSL) in particular (Efthimiou and Fotinea, 2007), which utilises the data of the GSL video corpus of ILSP, indicates that GSL utterances can be

analysed as surface realisations of recurrent underlying syntactic structures, in which head morphemes with well defined grammatical function are placed in standard positions in a string-like order (Efthimiou, 2008).

The theoretical linguistic study of classifiers builds upon and expands on previous work (Sutton-Spence and Woll, 1999; Berenz, 2002; Efthimiou et al., 2008), being especially concerned with the satisfactory treatment of the so-called Classifier Predicates (CPs) of SLs within theoretical-linguistic frameworks of analysis (Cogill-Koez, 2000), which have historically evolved in parallel with the study of spoken languages.

A problem posed by the second fundamental Saussurean principle of linguistic analysis is that of the Arbitrariness of the Sign: Classifier Predicates utilise standard handshapes (the so-called “classifiers”) to directly denote certain salient geometrical properties of the referents referred to by the nominal arguments of two- and three-place SL predicates. In other words, the signal (the handshape) denoting the signified concept (the geometrical property of the referent) is highly motivated (to a certain degree, non-arbitrary) in terms of physical resemblance. The element of iconicity is very strongly present in the signals realising CPs, and, indeed, far more strongly so than in the signals realising the nominals which refer to the real-world objects and whose relationship is denoted through the semantics of the predicator. This latter fact has led certain linguists to characterise SL signs corresponding to concepts which a spoken language would signify by a concrete noun as “frozen” (Cogill-Koez, 2000).

To complicate matters further, the direction of movement within signing space of classifier-handshapes realising/participating in CPs is a direct spatial metaphor of the physical relation between the referents denoted by the nominals realising the arguments of the predicate. More specifically, the position and the direction of movement of the classifier-handshapes with respect to the position of the signer’s body is a direct spatial metaphor

denoting the θ -roles (e.g. agent, recipient, location, etc) performed by the nominal arguments.

Theoretical analysis of the linguistic data available in the classifier elicitation corpus (2.1 above), supports formulation of a proposal for classifier behaviour which differentiates among three distinguished major grammar functions (Efthimiou and Fotinea, to appear).

Based on the key role of classifiers to behave as semantic markers which create semantic classes of objects, we propose an analysis of CPs which utilises classifier morphemes in three distinct ways:

- i) Classifiers create new lexicon items: Classifier affixation adds specific semantic properties to an entity, making it part of the semantic class this specific classifier identifies. In GSL, lemmas like 'GLASS', 'AIRPLANE', 'WALK', 'TABLE' etc., or handshapes like C, B, etc, may undertake classifier function. This is especially productive in the case of concrete object linguistic representations, e.g. the sign 'PENCIL' utilises classifier Δ (delta), the sign 'BOTTLE' utilises classifier C, the sign 'FIELD' utilises classifier 5, etc.
- ii) Classifiers add qualitative/quantitative values: Classifiers function as modifiers adding qualitative/quantitative values to syntactic heads or maximal phrases (i.e. boxes of different volume, pipes of different size, raising objects of different weight).
- iii) Classifiers serve co-indexing: In sign utterances, classifiers may be used as pronominal elements, where co-indexing obligatorily involves an expanded set of agreement features which, apart from the standard features "Number" and "Gender", also includes the feature "Semantic Class". Indicative examples of such formations are sign phrases elicited via stimuli as those presented in pictures c1, c2 and c3 of figure 1.

5. Future research perspective

The here reported research work provided a basis for a unified analysis of classifier functions in GSL. Next steps include verification of our hypotheses by elicitation of further related data but also a more concrete classification scheme. With the existing categories and additional signing data we are opting to enrich our coding scheme with more examples and eventually limit the annotation categories to 20, so that each Classifier will be described with no more than 5-7 annotation categories.

This will facilitate the creation of an operational set of annotation categories for the description of classifiers, which will also enable implementation of an educational environment for the use of Classifiers in GSL.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135. The authors wish to thank K. Perifanos for his assistance in implementation of the video search tool.

7. References

- Berenz, N. (2002), Insights into person deixis". *Sign Language and Linguistics (SL&L)* 5:2, pp. 203--227.
- Cogill-Koez, D. (2000). Signed language classifier predicates: Linguistic structures or schematic visual representation?. In *Sign Language & Linguistics Journal* 3:2, pp. 153--207.
- Efthimiou, E. (2008). Processing cumulative morphology information in GSL: the case of pronominal reference in a three-dimensional morphological system. In Mozer, Charalambakis, Bakakou-Orfanou and Chila-Markopoulou (eds), pp. 114-128.
- Efthimiou, E., and Fotinea, S.-E.. The geometry of Semantics: The spectrum of Classifier functions in Greek Sign Language Grammar. *Journal of Greek Linguistics (JGL)* (to appear).
- Efthimiou, E., Fotinea, S.-E., and Sapountzaki, G. (2008). Feature-based natural language processing for GSL synthesis. *Sign Language and Linguistics Journal (SL&L)* 10:1. pp. 3--23.
- Efthimiou, E., and Fotinea, S.-E. (2007). GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In *Proceedings of 12th International Conference on Human-Computer Interaction: Universal Access in HCI*, Part I, HCII 2007, LNCS 4554, pp. 657--666.
- Efthimiou, E., Fotinea, S.-E., and Sapountzaki, G. (2006). Processing linguistic data for GSL structure representation. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Lexicographic matters and didactic scenarios*, Satellite Workshop to LREC-2006 Conference, May 28, pp. 49--54.
- Karpouzis, K., Caridakis, G., Fotinea, S.-E., and Efthimiou, E. (2007). Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture. *Computers and Education*, Elsevier, Volume 49, Issue 1, August 2007, pp. 54--74, electronically available since Sept 05.
- Sutton-Spence, R., and Woll, B. (1999). *The Linguistics of British Sign Language; an Introduction*. Cambridge University Press.

Dicta-SIGN - Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication

Eleni Efthimiou¹, Stavroula-Evita Fotinea¹, Thomas Hanke², John Glauert³, Richard Bowden⁴,
Annelies Braffort⁵, Christophe Collet⁶, Petros Maragos⁷, and François Goudenove⁸

¹Institute for Language and Speech Processing, ²Universität Hamburg, ³University of East Anglia, ⁴University of Surrey,

⁵LIMSI/CNRS, ⁶Université Paul Sabatier, ⁷National Technical University of Athens, ⁸WebSourd

E-mail: ¹{eleni_e, evita}@ilsp.gr, ²thomas.hanke@sign-lang.uni-hamburg.de, ³J.Glauert@uea.ac.uk,

⁴R.Bowden@surrey.ac.uk, ⁵annelies.braffort@limsi.fr, ⁶collet@irit.fr, ⁷maragos@cs.ntua.gr,

⁸francois.goudenove@websourd.org

Abstract

Here we present the components and objectives of Dicta-Sign, a three-year FP7 ICT project that aims to improve the state of web-based communication for Deaf people by allowing the use of sign language in various human-computer interaction scenarios. The project researches and develops recognition and synthesis engines for sign languages at a level of detail necessary for recognising and generating authentic signing. In this context, Dicta-Sign aims at developing several technologies demonstrated via a sign language-aware Web 2.0, combining work from the fields of sign language recognition, sign language animation via avatars, sign language linguistics, and machine translation, with the goal to allow Deaf users to make, edit, and review avatar-based sign language contributions online, similar to the way people nowadays make text-based contributions on the Web.

Dicta-Sign supports four European sign languages: Greek, British, German, and French Sign Language and differs from previous work in that it aims to integrate tightly recognition, animation, and machine translation. All these components are informed by appropriate linguistic models from the ground up, including lexical and grammar modelling, manual and non-manual features.

1. Rationale

The development of Web 2.0 technologies has made the WWW a place where people constantly interact with another, by posting information (e.g. blogs, discussion forums), modifying and enhancing other people's contributions (e.g. Wikipedia), and sharing information (e.g., Facebook, social news sites). Today's predominant human-computer interface, is relatively manageable for most Deaf people: The use of a language foreign to them is restricted to single words or short phrases. The graphical user interface, however, puts rather severe limitations on the complexity of the human-computer communication, and therefore it is expected that it will be replaced in many contexts by human language interaction. Obviously, a far better command of the interface language is required here than in graphical environments. Most Deaf people would therefore be excluded from this future form of human-computer communication unless the computer is able to communicate in sign language. Moreover, exclusion is already experienced with regard to interpersonal communication between Deaf individuals, given the current lack of translation tools to support SL-to-SL but also oral-to-SL and SL-to-oral applications. Sign language videos are not a viable alternative to text, for two reasons: Firstly, they are not anonymous – individuals making contributions can be recognized from the video and therefore limits those willing to contribute. Secondly, people cannot easily edit and add to a video that someone else has produced, so a Wikipedia-like web site in sign language is currently not possible. In order to make the Web 2.0 fully accessible to Deaf people, sign language contributions must be displayed by an animated avatar, which addresses both anonymisation and easy editing.

2. The Dicta-Sign project

Dicta-Sign is a project aimed at developing the technologies required for making sign language-based Web contributions possible, by providing an integrated framework for sign language recognition, animation, and language modelling. It targets four different European sign languages: British (BSL), German (DGS), Greek (GSL) and French (LSF), and develops three proof-of-concept prototypes: a search-by-example sign language dictionary, a sign language-to-sign language translator, and a sign language-based Wiki.

A key aspect of the Dicta-Sign project is the creation of parallel corpora with detailed annotations in the four above-mentioned signed languages. These not only greatly aid the development of language models for both recognition and animation, but also allow for the direct spatio-temporal alignment of equivalent utterances across the four languages, which is useful for creating machine translation algorithms in a sign language-to-sign language translator.

3. Objectives

One of the main objectives of Dicta-Sign is to develop an integrated framework that allows contributions in the four sign languages of the project. Users make their contributions via webcams. These are recognized by the sign language recognition component and converted into a linguistically informed internal representation, which is used to animate the contribution with an avatar, and to translate it into the other respective three sign languages. Other objectives include the development of the world's first parallel multi-lingual corpus of annotated sign language data; the development of advanced sign language annotation tools that integrate recognition,

translation, and animation; the provision of large cross-lingual sign language dictionaries; and the advancement of the state of the art in computer vision and sign language recognition, sign language generation, sign language linguistic modelling and sign language translation.

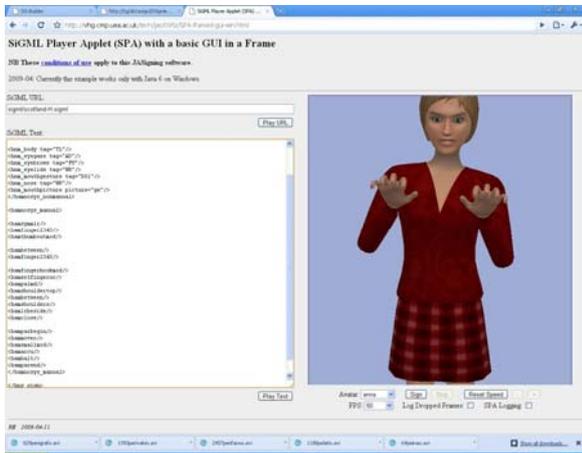


Figure 1: Avatar frontal view demonstrating incorporation of non-manuals in the sign synthesis engine: eyes, eyebrows, mouth and body posture participate in sign articulation

Dicta-Sign is working closely with the Deaf communities in the countries of the project partners throughout the lifecycle of the project to ensure that its goals are met and to evaluate user acceptance.

4. Research Domains

Research activities within Dicta-Sign expand from sign language recognition to sign synthesis and animation, linguistic modelling and development of annotation tools.

4.1 Sign Language Recognition

Despite intensive research efforts, the current state of the art in sign language recognition leaves much to be desired. Problems include a lack of robustness, particularly when low-resolution webcams are used, and difficulties with incorporating results from linguistic research into recognition systems. Moreover, because signed languages exhibit inherently parallel phenomena, the fusion of information from multiple modalities, such as the hands and the face, is of paramount importance. To date, however, relatively little research exists on this problem (Ong & Ranganath, 2005). The features that serve as input to the recognition system comprise a mix of measurements obtained by statistical methods, and geometrical characterisations of the signer's body parts. In order to make the feature extraction process robust even when the image comes from commodity webcams, the computer vision algorithms need to operate on multiple scales. Moreover, the basic feature extraction processes need to be combined with statistical and learning-based methods, such as active appearance models for facial expression tracking (Cootes et al., 2001; Papandreou & Maragos, 2007).

Sign language is inherently multimodal: both hands move

in parallel, while the face and body exhibit grammatical and prosodic information (Neidle et al., 2000). Hence, sign language recognition must deal with the problem of fusing multiple channels of information.

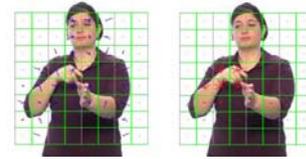


Figure 2: Figure showing HOG/HOF features (Histogram of Oriented Gradients, Histogram of Optical Flow) from a single frame of Sign Footage

Given the current state of the art in sign language recognition, one cannot expect the system to recognize the full range of expressiveness in signed languages. We deal with this limitation in two ways: First, the prototype application is domain-specific, with a restricted vocabulary of no more than 1500 signs. Second, the system employs a dictation-style interface (hence the name "Dicta-Sign"), where the user is presented with the closest-matching alternatives if a sign is not recognized reliably.

The output of the recognition component is converted into a linguistically informed representation that is used by the synthesis and language modelling components, respectively.

4.2 Synthesis and Animation

In the Dicta-Sign project, the internal representation of sign language phrases is realized via SiGML (Elliott et al., 2000), a Signing Gesture Markup Language to support sign language-based HCI, as well as sign generation. The SiGML notation allows sign language sequences to be defined in a form suitable for execution by a virtual human, or avatar, on a computer screen. The most important technical influence on the SiGML definition is HamNoSys, the Hamburg Notation System (Hanke, 2004), a well-established transcription system for sign languages. The SiGML notation incorporates the HamNoSys phonetic model, and hence SiGML can represent signing expressed in any sign language.

One of the most difficult problems in sign synthesis is converting a linguistic description of the signed utterance into a smooth animation via inverse kinematics, with proper positioning of the hands in contact with the body, and generating realistic prosodic features, such as appropriate visual stress. To this end, the Dicta-Sign corpus, does not only encompass phonetic and grammatical information, but also prosodic information. Together with the features derived from the visual tracking and recognition component, this allows for greatly increased realism in the animations.

4.3 Linguistic Modelling

Linguistic modelling will develop a coherent model from the phonetic up to the semantic level of language representation, envisaged to be language-independent in

most aspects. Dicta-Sign aims to extend modelling capabilities toward a common representation of sign language grammar and the lexicon -or alternatively two coherent representations- to accommodate both sign language recognition and synthesis. Overall, this represents a major advance over previous work, since language modelling has been largely neglected particularly in the recognition field.

4.4 Annotation Tools

Although some tools exist for specifically processing signed languages, such as iLex, none of these tools currently provide any kind of automated tagging, so the annotation process is completely manual.

An experimental version of the AnCoLin annotation system allows some image processing tasks to be initiated from within the annotation environment and to compare the results with the original video (Braffort et al., 2004; Gianni et al., 2007). It also connects to a 3D model of the signing space, but still lacks a coherent integration into the annotation workflow. It is expected that one of the major outcomes of the Dicta-Sign project will be greatly improved annotation tools, with image processing and recognition integrated into the annotation workflow. Their long term utility can be judged by the uptake by other sign language researchers.

4.5 Sign Language Corpora

A substantial corpus is needed to drive automatic recognition and generation, so as to obtain sufficient data for training and language representation. The quality and availability of sign language corpora has improved greatly in the past few years (Efthimiou & Fotinea, 2007; Neidle & Sclaroff, 2002). Yet, to date, multi-lingual sign language research has been hampered by the lack of sufficiently large parallel sign language corpora. One of the most important goals of Dicta-Sign is to collect the world's first large parallel corpus across four signed languages (Greek, British, German, and French).

This corpus will be annotated, showcase best practices for sign language annotations, and be made available to the public.



Figure 3: Two handed sign articulation by neutral body posture

5. Expected Outcomes

Expected outcomes of the project expand to both

prototype systems of SL technologies and SL resources, and include:

- A parallel multi-lingual corpus for four national sign languages – German, British, French and Greek (DGS, BSL, LSF and GSL respectively) – of a minimum of three hours signing in each language,
- A substantial dictionary of at least 1500 signs for each represented sign language,
- A continuous sign language recognition system that achieves significant improvement in terms of coverage and accuracy of sign recognition in comparison with current technology; furthermore this system will research the novel directions of multimodal sign fusion and signer adaptation,
- A language generation and synthesis component, covering in detail the role of manual, non-manual and placement within signing space,
- Annotation tools which incorporate these technologies providing access to the corpus and whose long term utility can be judged by the up-take by other sign language researchers,
- Three bidirectional integrated prototype systems which show the utility of the system components beyond the annotation tools application,
- A showcase demonstrator which exhibits how integration of the different components can support user communication needs.

6. Proof-of-concept Prototypes and Project Demonstrator

Three proof-of-concept prototypes will be implemented and evaluated within Dicta-Sign:

- A Search-by-Example system will integrate sign recognition for isolated signs with interfaces for searching an existing lexical database.
- An SL-to-SL translation prototype will pioneer a controlled-vocabulary sign language-to-sign language translation on the basis of the parallel language resources developed within the project.
- A Sign-Wiki will be developed providing the same service as a traditional Wiki but using sign language.

As a showcase of the different technologies developed within Dicta-Sign, an SL-to-SL terminology translator will be developed to serve as project demonstrator.

7. Conclusion

Today, still living in the atmosphere of the “European Year of Equal Opportunities for All,” it is important that drastic measures are taken to prevent new barriers from arising, as new forms of communication establish their role in the society at large. Dicta-Sign will be a key technology to promote sign language communication, and to provide Web 2.0 services and other HCI technologies to Deaf sign language users, an important linguistic minority in Europe, so far excluded from these new developments.



Figure 4: Examples of coarse pose estimation during Signing



Figure 5: Motion estimation and segmentation

As the field of sign language technology is still very young, it is beyond the scope of a three-year project to catch up completely with mainstream language technology, and to deliver end-user products. Nevertheless, Dicta-Sign is poised to advance significantly the enabling technologies by a multidisciplinary approach, and to come close enough to let designers of future natural language systems fully take sign languages into account.

8. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

9. References

- Braffort, A., Choisier, A., Collet, C. (2004). Toward an annotation software for video of Sign Language, including image processing tools and signing space modelling. In *Proceedings LREC 2004*.
- Cootes, T. F. and Edwards, G. J. and Taylor, C. J. (2001). Active Appearance Models. In *IEEE Trans. PAMI*, vol.23, no.6, pp. 681--685.
- Efthimiou, E. & Fotinea, S-E. (2007). GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI. In *Proceedings of 12th International Conference on Human-Computer Interaction: Universal Access in HCI*, Part I, HCI 2007, LNCS 4554, pp. 657--666.
- Elliott, R., Glauert, J.R.W., Kennaway, J.R., and Marshall, I. (2000). Development of Language Processing Support for the Visicast Project. In *Proceedings ASSETS 2000 4th International ACM SIGCAPH Conference on Assistive Technologies*, Washington DC, USA, 2000.
- Gianni, F., Collet, C., Dalle, P. (2007). Robust tracking for processing of videos of communication's gestures. In *Proceedings International Workshop on Gesture in Human-Computer Interaction and Simulation (GW 2007)*, Lisbon, Portugal, May 2007.
- Hanke, T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In *Proceedings LREC 2004, Workshop proceedings: Representation and processing of sign languages*. Paris: ELRA, 2004, pp. 1--6.
- Marshall, I., Sáfár, E. (2005). "Grammar Development for Sign Language Avatar-Based Synthesis", In *Proceedings HCI 2005, 11th International Conference on Human Computer Interaction (CD-ROM)*, Las Vegas, USA, July 2005.
- Neidle, C., Kegl, J., MacLaughlin, D., Bahan, B. and Lee, R.G. (2000). *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. Cambridge, MA. MIT Press.
- Neidle, C. and Sclaroff, S. (2002). Data collected at the National Center for Sign Language and Gesture Resources, Boston University. Available online at <http://www.bu.edu/asllrp/ncslgr.html>
- Ong, A.C.W. and Ranganath S. (2005). Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning. In *IEEE Trans. PAMI*, 27(6): pp. 873--891, 2005.
- Papandreou, G. and Maragos, P. (2007). "Multigrid Geometric Active Contour Models. In *IEEE Trans. Image Processing*, vol.16(1), pp. 229--240.

Towards the Integration of Synthetic SL Animation with Avatars into Corpus Annotation Tools

Ralph Elliott[†], Javier Bueno[‡], Richard Kennaway[†], John Glauert[†]

[†]School of Computing Sciences, UEA Norwich, Norwich NR4 7TJ, UK

[‡]Departamento de Ciencias de la Computación, Universidad de Alcalá, Alcalá de Henares, Spain

R.Elliott@uea.ac.uk, fjavier.bueno@uah.es, R.Kennaway@uea.ac.uk, J.Glauert@uea.ac.uk

Abstract

We outline the main features of our synthetic virtual human sign language system, JASigning. We describe how we have extended its input notation, SiGML, to allow explicit control of performance time, and we describe our initial steps on the path to integrating virtual human sign language performance into annotation tools, where it may be compared with video depicting the corresponding real human performance.

1. Introduction¹

JASigning is the current incarnation of our earlier synthetic virtual human signing system, SiGMLSigning. Like its predecessor, the system uses SiGML as its input notation. In this paper we start with a brief overview of the system before going on to describe our recent work in comparing virtual human sign language performance with real human signing as recorded in video sequences. We describe the introduction of explicit timing features into SiGML, and the way this can be exploited when making the comparison between real and virtual human signing. Finally we describe our initial moves towards the integration of virtual human signing into sign language annotation tools, and consider briefly the benefits of this integration.

2. Background

2.1. The JASigning System

JASigning (Java Avatar Signing) is a synthetic sign language animation system. In terms of its capabilities, JASigning is very similar to the SiGMLSigning system that we developed a few years ago in the ViSiCAST and eSIGN projects (Elliott et al., 2004; Elliott et al., 2007).

Thus JASigning supports both desktop and Web applications (Figure 1) that allow the user to have a virtual human, or avatar, perform a sign language sequence described in the SiGML (Signing Gesture Markup Language) notation. The system operates in real-time, so the SiGML sequence performed by the avatar at any point in time may be selected, or even generated dynamically, in response to user interaction. The most prominent difference between JASigning and SiGMLSigning is that the earlier system could run only on Windows computer systems, whereas JASigning, whose avatar software is implemented in Java, can be deployed on multiple platforms. It is currently available on both Windows and Mac OS X systems.

2.2. The SiGML Notation

As we have said, the input notation for any JASigning application is SiGML (Elliott et al., 2004; Elliott et al., 2007),

¹We acknowledge with gratitude that the work described here has been partially funded under the European Union's 7th Framework Programme, through the Dicta-Sign project (grant 231135).

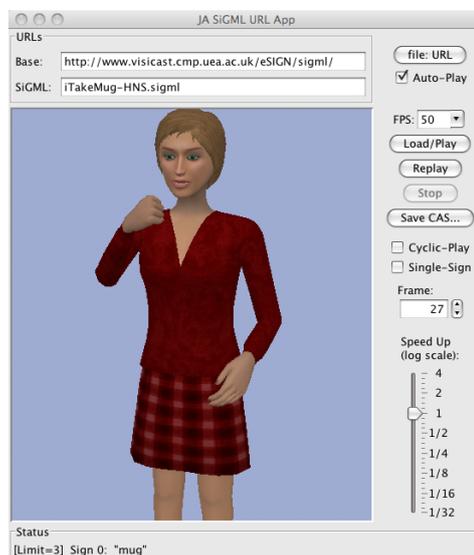


Figure 1: SiGML URL Player Application

an XML application which is based closely on HamNoSys (Hamburg Notation System) (Prillwitz et al., 1989; Hanke, 2004), and which is thus a vehicle for sign language description at the phonetic level.

The basic notions in the HamNoSys/SiGML model are those of posture and movement (transition). The manual component of a posture is characterised by its handshape, its spatial orientation, and its location in signing space — these features being specified for the dominant hand only in a single-handed sign, or for both hands in a two-handed sign. A basic movement consists of a change in some aspect of posture. These changes may be combined either concurrently, where that makes physical sense, or in sequence.

Historically, HamNoSys focused predominantly on the definition of the manual features of sign language performance, but its current version, HamNoSys 4 defines a comparatively rich repertoire of nonmanual features on different tiers, corresponding to distinct articulators such as body, eyes and mouth. SiGML follows HamNoSys 4 in including this repertoire of nonmanual features.

In many ways the HamNoSys SiGML model as just described resembles the phonetic model for sign languages of Liddell and Johnson (Liddell and Johnson, 1989), although there are some significant points of difference.

A SiGML document is structured as a sequence of individual signs. The notation allows sign language sequences to be represented in several distinct forms, of which the two most important are:

- **HNS-SiGML** In essence this is simply HamNoSys dressed in XML form, one element per symbol.
- **Gestural SiGML** This contains the same information as an HNS-SiGML or HamNoSys definition (in fact, potentially a slightly generalised version of this information), but in a more explicitly structured form, comparable to that of an abstract syntax tree for the corresponding HamNoSys or HNS-SiGML definition.

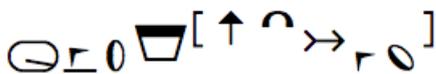


Figure 2: Manual HamNoSys Sign – “mug” in BSL

```
<hamgestural_sign gloss="mug">
  <sign_nonmanual>
    <mouthing_tier>
      <mouth_picture picture="mVg"/>
    </mouthing_tier>
  </sign_nonmanual>
  <sign_manual>
    <handconfig handshape="fist"
      thumbpos="across"
      extfdir="ol" palmor="l"/>
    <location_bodyarm
      location="shoulders"/>
    <par_motion>
      <directedmotion
        direction="u" curve="u"/>
      <tgt_motion>
        <change posture/>
        <handconfig
          extfdir="ul" palmor="dl"/>
      </tgt_motion>
    </par_motion>
  </sign_manual>
</hamgestural_sign>
```

Figure 3: Gestural-SiGML Sign – “mug” in BSL

In Figure 2 we show the HamNoSys for the manual component of the BSL sign “mug”, a snapshot of which is shown in Figure 1. The first three symbols describe the handshape and orientation, and the fourth the location (shoulder-level), for the initial posture; the remaining symbols specify a composite movement from this posture. Figure 3 shows the Gestural SiGML form of this sign. The motion from the initial posture, once attained, is a composite of two basic motions performed in parallel, that is, concurrently: an

upwards curved motion of the dominant hand, and a change of hand orientation. Together these motions function iconically, tilting the hand (whose shape itself functions iconically to represent a mug) towards the signer’s mouth. In the HamNoSys (and HNS-SiGML) forms the fact that these motions are performed concurrently with one another is indicated by the pair of square bracket symbols, whereas in the Gestural SiGML form the motion structure is directly reflected in the XML element structure, in which a `par_motion` element has a child element for each of the two component motions — the `directedmotion` and the `tgt_motion` (targetted motion).

SiGML can effectively be regarded as a kind of programming notation for the avatar: in principle any sign language utterance can be described in SiGML, as in HamNoSys; hence it can be performed by an avatar in the JASigning system.

2.3. Organisation of the JASigning Software

A signing avatar in the JASigning system is based on conventional 3D computer animation techniques. These techniques are augmented with additional data files defining those characteristics of the avatar that are needed for sign language performance — described in a companion paper (Jennings et al., 2010) — and with a software module, Animgen (Kennaway et al., 2007), whose function is to generate a sequence of animation frames, each defining an instantaneous posture for a specific avatar. Animgen does this given two inputs: the (avatar independent) SiGML description of the required sign language sequence, and the dataset describing the avatar for which the animation is required.

3. Working towards Integration with Annotation Tools

The synthetic sign language animation system is certainly still capable of further refinement and improvement, but it has reached a stage of maturity at which it is feasible to consider how it might be integrated into sign language annotation tools such as ELAN (Hellwig et al., 2009) and ILex (Hanke, 2004), and what the benefits of doing this might be. We outline here our recent activities in this area, the first of which involves an extension to the SiGML notation and its implementation.

3.1. Introduction of Explicit Timing into SiGML

The timing model for sign language performance used by JASigning’s animation generation module, Animgen, can be described as follows. Each basic movement is assigned a supposedly “natural” duration. This is done by means of one of the avatar-specific configuration data files described in the companion paper (Jennings et al., 2010). Hence, for a given avatar it is possible to vary these individual duration values relative to one another, and also to vary some or all of these configuration parameter values from one avatar to another. In addition, a configuration parameter determines the “natural” value for the movement to the initial posture of a sign. Once fixed, these duration values for basic movements determine those for composite movements. In the case of a sequence of movements, the duration of the sequence is simply the sum of the individual component du-

ration values. For a parallel combination of movements the overall duration is the longest of the individual component durations, the other component durations being extended to that maximum value.

We have recently extended the SiGML notation, and its implementation in Animgen, to allow explicit timing characteristics to be attached both to any individual motion, whether basic or composite, and also to an entire sign.

This is done by means of an additional pair of attributes, each with a floating point value, either or both of which may be attached to any relevant Gestural SiGML component:

- `duration`, measured in seconds, whose default value is the “natural” duration value, as described above.
- `timescale`, a slow-down factor, whose default value is 1.0.

(There is a third attribute, `speed`, whose effects are identical to those given by using the `timescale` attribute with the reciprocal value, so we omit it from the following discussion.) For any motion, if its (explicit or default) `duration` and `timescale` values are, respectively, d and t , then the duration value assigned to it, a , is given by the formula:

$$a = d * t$$

(according to which, the default duration value is indeed the “natural” one).

Whenever a composite motion, including an entire sign, is explicitly given a non-standard duration value in this way, that value is propagated down the motion structure as follows. Any increase or decrease in the duration of a composite motion is propagated to each of its components in proportion to the relative durations assigned to them prior to this adjustment. Any increase or decrease in the duration of a parallel motion is applied to each of its constituent motions (which in some cases may simply be a matter of undoing, to some degree, a previously applied extension). If any constituent motion is itself composite, its new duration value is propagated recursively to its components.

3.2. Comparing Virtual and Real Sign Language Performance

Our first activity in this area consisted of an investigation of the fidelity with which the signing avatar system could reproduce some Spanish Sign Language (LSE) sequences for which video material was already available. This was partly a matter of considering the basic quality of the animation produced from a HamNoSys or SiGML transcript, and partly a matter of determining the extent to which it is possible to improve the fidelity of the animation by adjusting the SiGML transcript, usually by adding more explicit detail relating to certain aspects of the original human performance.

To compare the results with the original it is useful to have video of the real and the virtual human performance side by side. This can be achieved by converting the animation system output to a video file, which can then be imported into an annotation tool. We have done this using ELAN 4.

An important issue for the comparison is that of synchronization, or the lack of it, between the real and the virtual animation. Using the SiGML enhancements for explicit timing control just described, it is relatively simple to align the two performances temporally, as is shown in Figure 4. So far we have pursued this only to the point of aligning sign boundaries, but in principle it is possible also to align individual movement phases within signs.

More recently, we have done some work with the ILex sign language corpus annotation tool (Hanke, 2004). ILex is able to export an annotation transcript, which includes segmentation and timing data, as well as a HamNoSys transcription of each sign. From this transcript we have been able to derive (almost) automatically a SiGML description of that sequence. When played by a signing avatar, the avatar performance exhibits some variations from that of the human signer in the video accompanying the transcript. In particular, as in the case of the LSE sequences described above, there are significant variations in the timing of the two performances.

Using the timing data from the ILex transcript, together with the new explicit timing attributes in SiGML, we have also been able to generate automatically a modified SiGML description of the sequence in which each sign is temporally aligned with its counterpart the original human performance. Thus from the exported ILex transcript we are able to produce a synthetic avatar performance – either in our avatar player, or exported from it as a video clip – which is temporally aligned, sign by sign, with the human performance.

A cursory comparison of the two performances gives rise to a couple of observations:

- The avatar makes some rather violent elbow movements, indicating scope for possible improvement of the generated animation.
- There are some variations in handshape and/or orientation, suggesting in some cases that the HamNoSys annotation may not be entirely accurate.

4. Conclusion

We have described the basic features of the JASigning synthetic signing system and the SiGML notation which is used to drive it. We have also described the introduction of explicit timing into SiGML and its implementation, and our moves towards the incorporation of virtual human signing into annotation tools, where it can be compared in detail with real human signing.

As yet, within an annotation tool (ELAN) we have augmented the original annotated video with the corresponding synthetic performance only in video form, but there is clearly no obstacle in principle to quite tight and interactive integration into an annotation tool of the process of generating and displaying synthetic sign language performance. On the basis of our experience to date, we can envisage several uses for such a scheme. As we have already seen, it can be used evaluate and to improve the quality of our synthetic sign language generation techniques.

Conversely, the capacity to get immediate feedback in the form of a synthetic animation provides a means of verifying

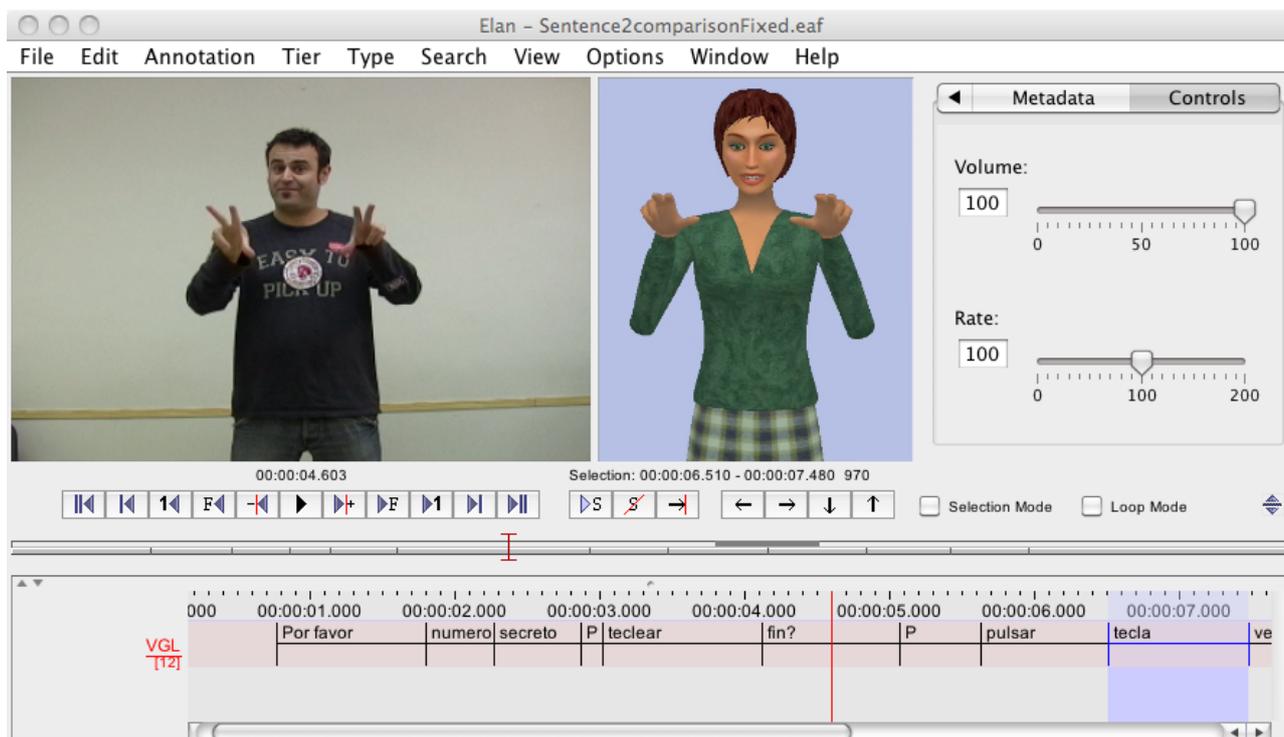


Figure 4: ELAN window with video of real and virtual human signers

the accuracy and quality of a HamNoSys transcription as soon as it has been generated. This can be useful both in the context of corpus collection and transcription, as well as in the context of signed content creation using a virtual human.

From the point of view of the kind of sign language study that annotation tools are intended to facilitate and support, the ability to compare and contrast virtual and real human sign language performance in great detail has the potential to assist in exploring more substantial questions in sign language modelling. For example, when confronted by variations between different performances of the same sign language sequence it is possible to ask whether these variations are linguistic in character, or whether they are matters of individual style or mood, whether they are peculiar to the particular utterance or part of a more persistent pattern. By taking our work further and fully integrating a synthetically signing avatar into an annotation tool, we can envisage a situation where it would be possible dynamically to modify some of the avatar's configuration parameters, for example those characterising its signing space, and exploring the way such variations cause the synthetic performance to align with or deviate from the original human performance. Experiments of this kind could help in leading to a richer characterisation — and hence annotation — of the original human sign language performance.

5. References

- R. Elliott, J.R.W. Glauert, V. Jennings, and J.R. Kennaway. 2004. An overview of the sigml notation and sigmlsigning software system. In O. Streiter and C. Vettori, editors, *Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 98–104, Lisbon, Portugal.
- R. Elliott, J.R.W. Glauert, R. Kennaway, I. Marshall, and E. Safar. 2007. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Universal Access in the Information Society*, 6(4):375–391.
- T. Hanke. 2004. Hamnosys—representing sign language data in language resources and language processing contexts. In O. Streiter and C. Vettori, editors, *Fourth International Conference on Language Resources and Evaluation, LREC 2004*, pages 1–6, Lisbon, Portugal.
- Birgit Hellwig, Dieter Van Uytvanck, and Micha Hulsbosch. 2009. Elan - linguistic annotator, version 3.8.
- V. Jennings, J.R. Kennaway, J.R.W. Glauert, and R. Elliott. 2010. Requirements for a signing avatar. In T. Hanke, editor, *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valletta, Malta.
- J.R. Kennaway, J.R.W. Glauert, and I. Zwitterlood. 2007. Providing signed content in the internet by synthesized animation. *ACM Transactions on Computer Human Interaction*, 14, 3(15):1–29.
- S.K. Liddell and R.E. Johnson. 1989. *American Sign Language : The Phonological Base*. Linstok Press.
- S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. 1989. *Hamburg Notation System for Sign Languages—An Introductory Guide*. International Studies on Sign Language and the Communication of the Deaf. IDGS, University of Hamburg.

Combining constraint-based models for Sign Language synthesis

Michael Filhol, Maxime Delorme, Annelies Braffort

LIMSI/CNRS

B.P. 133, 91 403 Orsay Cedex

E-mail: michael.filhol@limsi.fr, maxime.delorme@limsi.fr, annelies.braffort@limsi.fr

Abstract

The framework is that of Sign Language synthesis by virtual signers. In this paper, we present a sign generation system using a variety of input layers, separated on two sides: an anatomical side and a linguistic side. In a first part we suggest a way of implementing the flexibility required by Sign Languages into the system by using combinations of necessary and sufficient constraints. The anatomical side of the input specifies all morphological and articulatory constraints that model the behaviour of a human skeleton, while the linguistic input specifies language constraints (lexical, grammatical, iconic...) that must be applied to the signer's body to utter the correct sign sequence. A second part explains how to combine all these parts of the input in a conjunction of constraints for each time frame of the animation. A point is made that conflicting constraints may be given and need be prioritised in order still to decide on acceptable solutions. A first idea of a global priority order is given to illustrate this issue.

1. Introduction & Context

Sign Languages (SLs) are the most natural way for the Deaf to communicate. Deaf people not all being comfortable with reading text, and for them to access everyday's information, we choose to combine audio information systems like station announcements with SL displays on screens. Those displays could play videos of people signing complete utterances but the nature of the information (generally flexible gap sentences) prevents us from doing so. A more flexible way of displaying SL on a screen is the use of a 3d signing humanoid called virtual signer (VS). A VS can be animated by hand, requiring professional and talented graphists, or by automatic generation, which requires all sorts of models. Since SLs are natural languages, they have their own syntax and lexicon that need to be modelled. For the signed output to be natural and understandable by deaf people, we also need realistic models for the VS: skeleton models, animation models and skinning models. This paper introduces a system combining several input models for the generation of signs. Section 2 addresses the models used, advocating the use of constraint-based models to synthesize signs and animate the VS. Section 3 deals with the construction of the final animation, by explaining how all parts of the total input are combined.

2. Using constraints as input for sign generation

The goal is to animate the VS with linguistically structured gesture. To carry out the task, it is therefore natural to consider at least a linguistic and an anatomical influence on the body. In this section we give an overview of the approach used for linguistic modelling in the system, then we discuss the anatomical model.

2.1 Linguistic Constraints

The linguistic side of the system generates the input coming from language-ruled principles such as lexical sign specification, grammatical structure or prosody. We presently only have a model for lexical description, called Zebedee, the grammatical layers remaining work in progress.

As we stated above, naturalness of the output animations

is also a goal for the task, and the tremendous flexibility of Sign Language makes it very challenging in that respect. So far, systems generating SL from formal input (Hanke, 2002) have used phonetic descriptions like HamNoSys (Prillwitz, 1989) that specify body (in fact here, mainly hand) activity for each lexical unit (sign). Our recent work (Filhol, 2006) explains that due to the parametric structure of the approach, flexible values become rigid. In other words, in a signed sentence, every described sign results in one and only signed form, thus the flexibility of signs is not accounted for.

To provide as much flexibility as possible, our work at LIMSI has been focusing on the design of models based on sets of constraints that avoid both under- and over-specification of what needs to be uttered (Filhol, 2009).

The basic Zebedee structure of a sign is a sequence of timing units (see 'TU's on fig. 1) aligned on a timeline, where each unit specifies **everything** that is required in the period of time it covers—like a certain direction along which to align a bone or a point where to place a body site—and **only that**. In other words, a minimal conjunction of lexically intended articulatory constraints is given for each timing unit, thereby building a set of (lexically) necessary and sufficient constraints (NSCs). Then, at any moment when signing, anything left unconstrained can virtually be performed in any possible way.

The point of avoiding over-specification is to leave things open for additional constraints to be added if needed, for reasons like:

- *iconicity*: 'citation form' of lexical units are often modified according to their iconic features to fit a given context (Zebedee handles that well);
- *role shifts*: when impersonating a character with a certain body posture while uttering a sign, all unconstrained articulators can be used for the shift, leaving the lexically constrained ones for the sign;
- *grammatical reasons*: if not required otherwise by the lexicon, grammar may require that the body lean forward (e.g. a form of future in LSF), raise the eyebrows (e.g. neutral yes/no question), and so forth;

- *anatomical reasons*, which are discussed in the next section;
- etc.

Similarly, all these influences on the body are specified with as many and as few constraints as possible. They will then be combined, together with those coming from the morphology of the body.

2.2 Anatomical Constraints

Linguistic constraints are not sufficient to build a correct sign. Since one of the priorities of the generation is the realism of the final animation, we need to add a little more information. The generation of signs can be summed up as the construction of N frames in which the skeleton of the VS must be set in a particular posture. The overall generation can thus be seen as the generation of a succession of postures. For each posture, the linguistic model gives us information on how some parts of the body should be placed and oriented in space. Finding a correct posture from this information is called inverse kinematics (IK). IK problems are often under-specified problems leading to many solutions for a given input. For instance placing the wrist at a specific location in space raises an infinity of solutions (rotation of the elbow). Considering a set of possible solutions to one problem, the only element that will allow us to prefer a solution to another is the naturalness of the pose. We then add information about how realistic a posture is by informing the resolution system about the nature of the skeleton. These anatomical constraints are of three kinds:

- *joint limits* give the range of motion of each degree of freedom of the skeleton, avoiding impossible angles for the body;
- *angle probability* tells how often a specific angle of a degree of freedom is found. This measure is built from general purpose motion capture databases (Carnegie Mellon University) and a statistical analysis (Delorme, 2010);
- *biomechanical data* enhances the general quality of the posture for specific joints. This data is applied on small portions of the skeleton like the hands (Neff, 2006). Since biomechanical simulations are usually time consuming we prefer the use of pre-computed tables instead of running a real-time model.

All of these constraints apply to the skeleton and will not be subject to variation throughout the whole synthesis. Thus, in order to generate the animation, we consider on one side constraints coming from a linguistic point of view, that define what is mandatory for the sign or sentence. On the other side, we look at constraints that apply to the body and stay constant through time. We are now going to see how these two kinds of constraints interact in the generation system for sign synthesis.

3. Combination of constraints

Using all these linguistic and anatomical constraints allows us to reduce the number of possible solutions, and eventually choose one as the best posture for a given problem (i.e. one frame). Figure 1 illustrates the layers of constraints generated by the different models mentioned in section 2, and what we mean by conjunction of constraints for each time frame.

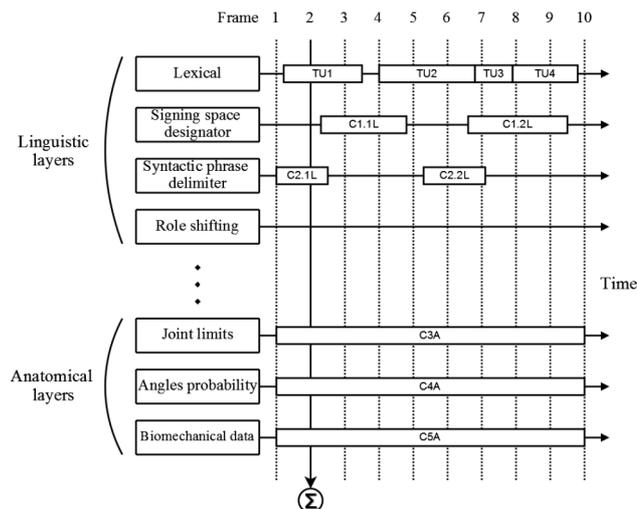


Figure 1: Resolution of multiple constraints through time

On the vertical axis we enumerate the layers of language (upper part of the list) and of anatomy (lower part) that may raise constraints on body articulations when signing. For instance, the purpose of the layer named "signing space designation" is to act on eye gaze and head (body articulators) as required in LSF to activate relevant parts of space or locate a new object by directing those articulators to the relevant points in space. The "syntactic phrase delimiter" will act on eyebrows and shoulders to mark topics in LSF, eyebrows for interrogatives, probably do some head shaking to emphasise negative clauses, etc. In the case of a dialog, "role shifting" will turn the body into the right direction to account for the alternating speakers. This layer will also use arms or hunch the back when impersonating characters with such distinctive markers.

We left the list of layers open as we imagine any number of them can be added to include more features, either additional language-specific rules, discourse prosody or indeed signing style, etc.

Theoretically, while the addition of constraints simply specifies the IK problem more (moving it away from under-specification), it also increases the risk for the problem to become over-specified (no solution).

A timeline is attached to each of these layers. In the diagram, time flows from left to right. When a layer generates a set of constraints over a period, they are represented by a white box on the timeline. As we said earlier, anatomical constraints remain constant in time, which is why the bottom lines have a box covering the whole animation without a change. At this point, it is clear that the set of constraints applying to the body at any moment in time is the conjunction of all constraints present on all layers at that moment.

Time is then broken in a sequence of frames to generate the output video. These time frames are shown across the drawing and numbered at the top, representing where to take snapshots of the timelines, each snapshot raising the set of constraints to combine hence a problem to solve for the time frame. On our example, frame no. 2 involves lexical constraints (from block TU1) and syntactic constraints (from C2.1L, say to mark the lexical sign as a

sentence topic), as well as all anatomical constraints (C4A, C5A and C6A). It bears no space designation constraint for instance, the first frame where these occur being frame no. 3, from block C1.1L.

While all constraints are given equal consideration, they may be processed in different stages of the synthesis. Constraints can be set to ask for contradictive or conflicting orders if two of them are located on the same parts of the skeleton. A good example of such conflict would be in French Sign Language (LSF) to sign "I know" while role shifting in a wolf character as illustrated in figure 2. To look more frightening, the signer frowns, hunches his back, raises his elbows and puts his hands (paws) forward. But to sign "I know", the signer needs to bring his strong hand to his forehead. So the system is given two orders regarding the right arm. There is no definite way of solving such conflicts since the priorities are sign-dependant. We chose to: first, give arbitrary priorities to the constraints, even if we know that this is not a really satisfactory solution; second, segment the skeleton into independent parts that will, to some extent, behave separately.



Figure 2: Left, "I know" in LSF; Right: the same sign while role-shifting as a wolf.

Here is an example of a simple priority scheme for constraints, based on the intuition of "what will work more often". This part of the work will of course need more investigation.

1. Joint limits are the absolute priority. We cannot have the VS make impossible angles.
2. Lexical constraints follow. They define as stated before what is absolutely necessary in the sign.
3. Grammatical layers add important information on the signs and must then be considered. Angle probabilities allow the system to choose in the resulting a set of solutions.
4. Finally, biomechanical data improves the configuration of unconstrained effectors (e.g. fingers) regardless of what has been previously computed.

The segmentation of the skeleton in five parts (see fig. 2) allows us to locate precisely which bone of the skeleton should be considered for a single problem. Thus a problem considering the right elbow will only involve the section "right arm", leaving the other parts free to be affected by different constraints. This might not be sufficient. For instance, a sign like [TREE] in French Sign Language needs the signer to place his weak hand on a specific location in space. Thus, considering only the hand from the wrist will fail. When no satisfactory

solution is found to a problem, the system tries again the resolution with a longer kinematic chain (i.e. a sequence of bone of the skeleton to move). In the precise case of [TREE], the system will consider the hand and the arm at the same time. If it still is not sufficient then the system will consider the complete kinematic chain including the hand, the arm and the spine (for instance signs needing to place the hand far from the body will lean the body forward to reach out further).

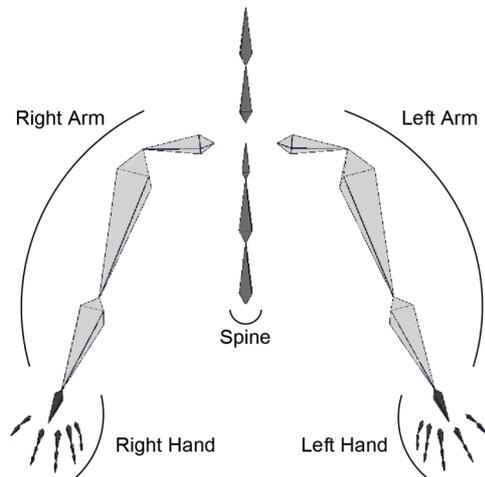


Figure 3 : Segmentation of the skeleton for progressive IK

The core of the resolution is the IK module. IK is a well-known problem in robotics (Lee, 1993) and animation (Komura, 2005). It consists in finding rotation angles for a kinematic chain to place an effector (e.g. the wrist, a finger, the elbow) at a specific location in space, or to orient it in a specific direction. The method we choose to solve IK problems is based on sequential Monte-Carlo simulations (Courty, 2008). This method is preferred to more common ones (Wang, 1991; Maciejewski, 1990) because of its very narrow connexions with probability distribution functions allowing us easily to include the anatomical constraints. The adaptation to our case works as follow:

1. We generate a certain number of random configurations for our skeleton. The range of the random angles is set to remain within the joint limits. Moreover, this generation follows the distribution functions of the angle probabilities to give more realistic results.
2. Every single solution is given a score depending on the quality of the result: in case of a placement the score depends of the distance between the effector and the target; in case of an orientation the score depends of the angle between the current orientation and the target orientation.
3. Each solution moves randomly around its current position trying to enhance its quality.
4. Biomechanical calibrations are made on the unprocessed parts of the skeleton to improve the overall posture.

The process iterates a limited number of times and stops if a good solution (given a threshold) is found. From this

process we extract the ten best results and assign scores to them, based on the angle probability tables. The more a configuration is found in the motion capture database, the higher its score. Finally we decide the most realistic solution is the one with the highest score and keep it as final result for the generation. This overall method is applied for each frame of the animation to generate.

4. Conclusion

We have presented a sign generation system based entirely on conjunction of constraints, coming from different layers of (for now, at least) language or anatomy. These constraints all apply to the skeleton of the VS but are synchronised differently in time according to the layer they belong to. The conjunction of all these constraints minimally specifies a posture for the skeleton at a specific time. As this can lead to conflicts, the constraints must be given relative priorities and a first tentative scheme was proposed. It should however be redefined from a precise analysis of which layers dominate the others, and indeed of whether they do constantly or in what way the scheme varies over time if not.

Such a system avoids too strong a separation between roles of articulators, e.g. dedicating the hands to the lexicon; the eyes to space activation and reference, and the torso to, say, role shifts. We separate the origins of the constraints in what we have called ‘layers’ of the system rather than what the constraints apply to. Now all layers may each act on all articulators.

Further work is needed to implement the system, as we currently have only anatomical and lexical constraints combined, but we hope this design brings to the field of sign generation more of, and an original approach to, the flexibility required by SLs.

5. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n°231135.

Some of the data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

6. References

- Carnegie Mellon University, Graphics Lab Motion Capture Database: <http://mocap.cs.cmu.edu/>
- Courty, N., Arnaud, E. (2008). Sequential Monte Carlo Inverse Kinematics v3. *INRIA Internal Report, RR-6426*, February 2008.
- Delorme, M. (2010). Sign Language Synthesis: Skeleton modelling for more realistic gestures. *SIGACCESS Newsletter*, February 2010.
- Filhol, M., Braffort, A. (2006). A Sequential approach to lexical sign description. In *LREC 2006 - Workshop on Sign Languages*, Genova, Italy.
- Filhol, M. (2008). Modèle descriptif des signes pour un traitement automatique des langues des signes, *PhD thesis*, Université Paris-11 (Paris sud), Orsay.
- Filhol, M. (2009). Zebedee: a lexical description model for Sign Language synthesis. *LIMSI internal report 2009-08*, Orsay.
- Hanke, T. et al. (2002). VISICAST deliverable D5-1: interface definitions. *VISICAST project report*.
- Komura, T., Ho, E.S.L, Lau, R.W.H. (2005). Animating reactive motion using momentum-based inverse kinematics. *Computed animation and virtual worlds*, vol. 16, pp. 213--223.
- Lee, S., Kim, S. (1993). Efficient inverse kinematics for serial connections of serial and parallel manipulators. In *Proceedings - RSJ/ICIRS, IEEE, Yokohama*, pp. 1635-1641.
- Maciejewski, A.A. (1990). Dealing with ill-conditioned equations of motion for articulated figures. *IEEE Computer Graphics Applications*, vol. 10, pp 233-242.
- Neff, M., Seidel, H-P (2006). Modeling Relaxed Hand Shape for Character Animation. *Articulated Motion and Deformable Objects (AMDO)*, vol. 4069 of LNCS, pp. 262--70.
- Prillwitz, S. et al (1989). HamNoSys version 2.0 - Hamburg Notation System for Sign Languages, an introductory guide. *Internation studies on Sign Language and communication of the Deaf*, vol. 5. Signum Press, Hamburg.
- Wang, L.T., Chen, C.C. (1991). A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. *IEEE Trans. Robotics Automation*, vol. 7, pp 489-499.

Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation

Jens Forster¹, Daniel Stein¹, Ellen Ormel², Onno Crasborn², and Hermann Ney¹

¹Human Language Technology and Pattern Recognition
RWTH Aachen University, Germany
{forster, stein, ney}@cs.rwth-aachen.de

²Department of Linguistics
Radboud University Nijmegen, Netherlands
{e.ormel, o.crasborn}@let.ru.nl

Abstract

We propose best practices for gloss annotation of sign languages taking into account the needs of data-driven approaches to recognition and translation of natural languages. Furthermore, we provide reference numbers for several technical aspects for the creation of new sign language data collections. Most available sign language data collections are of limited use to data-driven approaches, because they focus on rare sign language phenomena, or lack machine readable annotation schemes. Using a natural language processing point of view, we briefly discuss several sign language data collection, propose best practices for gloss annotation stemming from experience gained using two large scale sign language data collections, and derive reference numbers for several technical aspects from standard benchmark data collections for speech recognition and translation.

1. Introduction

Data-driven approaches to spoken language recognition and translation have seen great success over the last years. Common to all data-driven approaches is the need of large amounts of annotated data to learn reliable statistical models. In the case of natural languages, a data-driven system tries to learn individual statistical models for each phoneme respectively word requiring the system to see several utterances of a phoneme or word in the training data. Most sign language data collections have been created for linguistic research and as such tend to focus on rare phenomena. Both the focus on less frequent sign language phenomena and a low type-token ratio have so far limited the application of data-driven approaches to recognition and translation of sign languages.

Assuming the point of view of data-driven approaches, we briefly discuss the status of several sign language data collections in Section 2. and describe the needs of data-driven approaches to natural language processing in Section 3. Based on the status of the discussed sign language data collections, the needs of data-driven approaches, and experience gained in working with two large scale sign language data collections, we propose best practices for gloss annotation of sign language data collections. The practices proposed in Section 4. are designed for easy application to new and existing sign language data collection and allow for linguistic accurate annotation.

If a new sign language data collection is to be generated, the choice of the domain and some derived technical aspects like the targeted type-token-ratio and the vocabulary size are crucial variables that have a high impact on the performance of data-driven approaches. Based on existing data collection designed for speech recognition and translation, we provide reference numbers that can be used in the planning step for new data acquisition in Section 5.. The paper is concluded in Section 6.

2. Sign Language Data Collections

Although a full review of all available data collections is out of the scope of this work, almost all available data collections consist of annotated video material of various signers. The annotation has been typically conducted in glosses using specialized annotation tools such as ELAN¹, iLex (Hanke and Storz, 2008), or Signstream (Neidle et al., 2001). Gloss annotation assigns each sign the word from a spoken language that most appropriately describes the meaning of the sign. Besides the gloss annotation scheme, HamNoSys (Prillwitz et al., 1989) strives to describe signs on a phoneme-like level. All data collections discussed in this section have been annotated using glosses. Figure 1 shows example images taken from all data collections discussed in this work.

The RWTH-BOSTON (Dreuw et al., 2008) data collections are annotated subsets of data originally recorded at the Boston University. The annotations have been adjusted by RWTH Aachen University to fulfill the requirements of data-driven approaches. The data collections contain vocabulary sizes of up to 483 glosses, up to four different signers signing predefined sentences in front of a uniform background. Due to the small size of the data collections, and gray scale and color video recordings from lab environments, the RWTH-BOSTON data collections permit rapid development and testing of data-driven techniques for continuous sign language.

The ATIS (Bungeroth et al., 2008) data collection contains parallel annotation and videos for English, German, Irish sign language, German sign language, and South African sign language in the domain of the Air Travel Information System (ATIS). While the data collection can be used to build direct translation systems between different sign languages, the total size of only 600 parallel sentences is small in comparison to other sign language data collec-

¹<http://www.lat-mpi.eu/tools/elan>

tions. From a recognition point of view, the ATIS data collection contains challenging video recordings conditions including stark changes in illumination, cluttered office environments, and partial occlusions of the signer. In addition to the challenging recording conditions, the ATIS data collection shows for all included languages a singleton fraction of over 50%.

The European Cultural Heritage Online organization (ECHO)² published sign language data collections for Swedish sign language, British sign language, and sign language of the Netherlands (Crasborn et al., 2004). Although the data collection shows a high number of types, the chosen domain of fairy tales is challenging for data-driven approaches because of the intensive use of classifier signs.

Corpus NGT (Crasborn and Zwitterlood, 2008) is a large scale data collection for sign language of the Netherlands from several domains. Domains include fable stories, cartoon paraphrases, and discussions on sign language and Deaf issues. Especially the later two domains are interesting for data-driven approaches, because they allow for free discussions on topics with inherent limited vocabularies and hardly any classifier signs. Furthermore, sentence-aligned translations are currently created for the two discussion domains in the context of the EU funded SignSpeak project.

The SIGNUM data collection (von Agriss and Kraiss, 2008) has been specifically recorded for data-driven recognition of German sign language. The data collection contains over 700 predefined sentences signed by each of the 25 different native signers, and setups for signer dependent and signer independent recognition. The signers were asked to wear dark clothes and were recorded standing in front of a dark background.

Finally, the RWTH-PHOENIX data collection described by (Stein et al., 2010) contains German sign language for the domain weather forecast. The video material is recorded from broadcast news aired on the German television station Phoenix. Beside gloss annotation of the signs, translations into German are provided by a state-of-the-art speech recognition system for German. The chosen domain and employed annotation scheme are chosen with data-driven approaches in mind.

3. Needs of Statistical Recognition and Translation

Data-driven approaches to pattern recognition and model learning strive to learn a statistical model from the provided input data that best explains the input data in terms of a provided annotation. In the case of sign language recognition, the input data is a video stream showing a signing person with the annotation being the assigned gloss. For data driven translation, the input is a text in the source language e.g. glosses and the annotation is the corresponding text in the target language e.g. spoken language. Since data-driven approaches try to explain the input data in terms of statistical models, a system needs to collect several different examples of data labeled by the same annotation to incorporate the typical variance of the input data into the statistical

model. Generally speaking, the more examples collected for a given annotation the better becomes the resulting statistical model.

In most cases the raw input data is difficult to explain by a statistical model due to high variance. Therefore, features are extracted from the raw data that allow for better discrimination between different annotations. The process of feature extraction strongly depends on the modality of the input data. While translation systems apply e.g. morphological parsing to the input data, vision-based recognition systems normalize the illumination, extract oriented gradients, and track the hands of the signer. Robust feature extraction in the presences of changing illumination, motion blur, scale changes, partial occlusion, and cluttered backgrounds is difficult to achieve using state-of-the-art computer vision techniques. Sign languages are especially prone to motion blur because of fast moving hands and abrupt motion changes. To ease the burden of feature extraction in video streams, we propose to limit the variability of the video streams by using standardized recording settings and high definition cameras capturing more than 30 frames per second.

Besides the statistical model explaining the input data, recognition and translation systems employ an additional knowledge source called the language model. The language model is learned from the annotations and assigns a probability to a sequence of annotations e.g. glosses based on the seen annotation sequences. Since the language model is learned from annotations, the language model depends on the domain of the annotations.

4. Best Practices for Gloss Annotation

Every variation in the annotation of a sign, though clearly identifiable by a human reader, will be treated as a new token by the computer. Minor concerns in the variation include spelling, capitalization, and linguistic comments within the annotations. While the first two minor issues can be enforced by the application of specific annotation parsers, linguistic comments contain additional information that cannot be extracted from a raw video stream. We propose to generally store all linguistic comments in a separate annotation or if you use ELAN a separate annotation tier.

4.1. Dialectic Signing Variants

A major issue in sign language annotation is the question of how to deal with dialectic signing variants. Dialectic signing variants of a word e.g. "MONDAY" are typically annotated by the same gloss in sign language data collections. However, dialectic variants of signs differ strongly in their appearance. If dialectic variants are annotated using the same gloss, a data-driven recognition system will learn a single model that tries to explain all dialectic variants of the sign in question. Ideally, each dialectic variant is represented by a distinct stochastic model that explains only this particular dialectic signing variant. To be able to train such a dialect specific model from data, the dialectic variants of a sign need to be consistently annotated by distinct glosses. Therefore, we propose to enumerate dialectic variants by applying the number as a postfix to the parent

²<http://echo.mpiwg-berlin.mpg.de/home>



Figure 1: Example images from different sign language data collections (f.l.t.r.): ECHO, Corpus-NGT, RWTH-BOSTON, RWTH-PHOENIX, ATIS, and SIGNUM

gloss e.g. “MONDAY1”, “MONDAY2”, etc. This procedure has been applied in creating the RWTH-BOSTON data collections and is applied in the extension of the RWTH-PHOENIX and Corpus NGT data collections. In order to keep track of the numerous dialectic variants and to keep the annotation consistent, we propose to build a database containing video examples of dialectic variants of every gloss.

4.2. Homonyms and Synonyms

Related to the question of dialectic signing variants is the question of how to annotate homonyms and synonyms. Special to sign languages is the fact that there are true homonyms such as the sign for “DOCTOR” and “BATTERY” in sign language of the Netherlands and homonyms that share the same manual components but differ in mouthing. While true homonyms do not pose a problem to data-driven approaches as long as they there are consistently annotated by the same gloss and a list of true homonyms is provided to ease data-driven translation, the second class of homonyms, called *Umbrella-Glosses* in Corpus NGT, requires special care. An example of such an *Umbrella-Gloss* is “PROGRAMMA” which, depending on the mouthing, can mean rules or laws in sign language of the Netherlands. We propose to either split the annotation of the manual and non-manual parts of a sign into separate annotation files or tiers annotating e.g. “PROGRAMMA” for the manual part and “REGELS” for the non-manual part. As an alternative, we suggest to annotate an *Umbrella-Gloss* by its umbrella class followed by a delimiter and the actual realization of the umbrella. An example of the later approach is “PROGRAMMA:REGELS” and “PROGRAMMA:WETTEN” found in Corpus NGT. An advantage of the later approach is that a list of *Umbrella-Glosses* can be automatically generated from the annotation files.

In the case of synonyms, human annotators tend to use the meaning of a sign that is most appropriate in the context of the current sentence. By doing so, a synonym sign gets annotated by different glosses in one data collection effectively taking away observations from the model to be learned for the core meaning of this sign and biasing models for the synonym meanings. Consider for example the German signs for cathedral and carnival which are synonyms for Cologne and occur frequently in German broadcast news. We propose to use the glosses “CATHEDRAL” and “CARNIVAL” instead of Cologne and mark the intended meaning by an additional explicit postfix such as “-(syn:COLOGNE)”. Again we propose to generate a database containing video examples of synonyms.

4.3. Compound Glosses

Besides homonyms and synonyms, there exist several sequences of signs that need to be annotated by a single compound gloss to encompass its full meaning. An example is the gloss for *gebarentaal* (sign language in Dutch) that is composed of the sign “GEBAREN” followed by the sign for “TAAL” in sign language of the Netherlands. From a speech recognition point of view, the best procedure is to learn distinct models for “GEBAREN” and “TAAL” while from a translation point of view it is best to learn a model for “GEBARENTAAL”. To cope with this mismatch and facilitate accurate linguistic annotation, we propose to separate the glosses for “GEBAREN” and “TAAL” by a distinct delimiter such as \wedge and to add the compound gloss “GEBARENTAAL” as additional information. This leads to a notation like “GEBAREN \wedge TAAL:GEBARENTAAL” as it has been adopted for Corpus NGT. A similar notation will be used in the extended RWTH-PHOENIX data collection.

4.4. Finger Spelling

Finger spelling has an analog in word spelling for spoken languages. In the annotation of spoken languages spelled characters receive distinct annotations so that data-driven recognition and translation systems can learn distinct models for each spelled letter. For the sign language data collections discussed in Section 2., a sequence of finger spelled letters is often annotated as a single gloss such as “TREE” rendering it indistinguishable from a sign that does not employ finger spelling. Again, a data-driven system would try to learn a model for “TREE” although distinct models for each of the spelled letters would be more robust because the models are not only learned from the spelling of “TREE” but from all occurrences of finger spelled letters. Therefore, we propose to either use distinct gloss annotations such as “T R E E” or to prefix finger spelled sequences by a special delimiter such as “#”. The later solution has the benefit that the amount of finger spelling in a data collection can be inferred automatically, it is less cumbersome to annotate, and existing annotations can be easily adapted to the proposed scheme.

4.5. Incorporation

Incorporation of signs is a common feature of sign languages. Typically two signs e.g. the sign for five and the sign for month are fused into a new sign featuring aspects of both parent signs. Since an incorporated sign is neither of the parent signs, a data-driven recognition system has to consider it a distinct class to be modelled from the given data. In order to distinguish the parent sign forms and incorporated sign form and to still keep information on the

parent signs, we propose to build the gloss annotation of the incorporated sign by connecting the glosses for the two parent signs by a hyphen e.g. “5-MONTH”. This scheme has been employed in the RWTH-PHOENIX data collection and is successfully used in data-driven recognition.

4.6. Pointing and Referencing Signs

One of the strengths of visual languages is the possibility to refer to specific points in the signing space. Signers typically use pointing and referencing signs to convey temporal and causal concepts as well as relations between persons and objects. Except for self-referencing, the meaning of pointing and referencing signs is context dependent. The context of a referencing sign (e.g. the name of a person) can normally not be observed from the referencing sign itself. Since the context is known to the annotators, they typically use the context of a pointing or referencing sign to gloss such a sign. The information that a pointing or referencing sign has been used is lost. Therefore, it is difficult for a data-driven recognition or translation system to train robust models for pointing and referencing signs. Additionally, stochastic model used for the context of a pointing or referencing sign (e.g. the sign for “TREE”) is biased by the visual content of the referencing sign. To limit the effects of annotating a pointing or referencing sign by its current context, we propose to include the context of a pointing or referencing sign as an additional information to the used gloss for pointing or referencing. As an example consider the notation adopted for Corpus NGT where a pointing/referencing sign is annotated by the gloss “IX” regardless of the intended context or e.g. consider the notation adopted for the RWTH-PHOENIX database where additionally to the gloss “IX” information on the spot in the signing space and the intended meaning is attached to the signing gloss as e.g. “-(loc:A,tree)”.

4.7. Classifier Signs

A typical feature in signed languages are classifier signs capitalizing on the concept of free movement of the hands within the signing space. Classifier signs are non-lexicalized signs that show extreme variance in appearance and production. While it is already difficult for human experts to describe and annotate the exact meaning of a classifier sign, data-driven approaches are so far not able to cope with them. We propose to mark classifier signs by a special tag such as the @ sign or “<CLASSIFIER>” to be able to automatically extract all classifier signs from a data collection or to be able to create subsets of a data collection without classifier signs. Besides the information that a classifier sign has been used, it is desirable to add the perceived meaning of the classifier sign as additional information to the gloss marking. The proposed handling of classifier signs has been successfully used in our work with data from Corpus NGT.

4.8. Machine-Readability

Finally, all proposed practices for gloss annotation are useless to the natural language processing community if the annotation itself is not machine readable, consistent, accurate, and adequate. Machine readability is a prerequi-

site to automatic processing and parsing of large amounts of annotation data. This aspect includes the question of the used character encoding, preferably “UTF-8”, and the choice of gloss delimiters. We propose to separate glosses by spaces and to avoid spaces within glosses and attached additional information. Further, we suggest to put additional information behind the relevant gloss annotation e.g. “GEBAREN^TAAL:GEBARENTAAL” and to use specific delimiters such as e.g. “^”, “-”, and “:” for different constructs as e.g. compound glosses or incorporation. In most annotation scenarios there will arise special cases requiring a special mark or prefix such as e.g. “@” or “#” to be applied to a gloss annotation. In such special cases, we propose to use unique marks not used in the remaining glossing scheme. The benefit of adhering to the proposed procedure is that the resulting annotation scheme is machine readable and can be automatically checked for consistency w.r.t. the chosen annotation scheme.

4.9. Adequacy of Annotation

Adequate annotation is crucial to data-driven systems because a data-driven system can only learn from data what can actually be seen in the data. For example, in most sign languages a negation of a sign is only conveyed by shaking the head parallel to performing the manual components of the sign. If a sign language recognition system is based on the manual components it will not be able to recognize the negation of a sign because the negation is only visible in the non-manual part. We suggest to split the annotation of manual and non-manual components such as eye gaze, shoulder movements, and facial expressions into distinct annotation files or tiers and to limit the annotation for each modality to what can actually be seen in the data for the modality in question at the given time. The proposed procedure eases the process of building specific statistical models for each modality and reduces errors in the systems. For data-driven translation, the parallel annotation of the glosses in another sign language or spoken language should be adequate in the sense that the glosses are translated as literally as possible without aiming for fluency in the target language. As an example a heavy nodding of the head accompanying the gloss “YES” we propose to translate by “yes, very much” rather than by “yes, I think this is a very good idea!”.

5. New Data Collections

Independent of the chosen language, data collections of natural languages are hardly usable for data-driven approaches if the needs of data-driven approaches (cf. Section 3.) have not been taken into account when creating them. Using two small scale data collections for speech recognition and translation as references, we propose reference numbers for several technical aspects of sign language data collections.

Tables 1 and 2 show the statistics of small scale data collections used in speech recognition and translation. Although these data collections are by far bigger than anything we will see for several years to come in sign language data collections, they are among to the smallest data collections available for data-driven approaches to spoken language recognition and translation. The Verbmobil II corpus depicted in Table 1 contains spontaneous German

Table 1: Speech Recognition – Verbmobil II Corpus, German language, Domain of travel and booking

	Training	Evaluation
# sentences	36,015	1,081
# running words	701,000	14,000
vocab. size	10,157	–
audio data [h]	61.5	1.6

Table 2: Speech Translation – IWSLT 2005 Corpus, Chinese-English, Domain of travel and booking

	Training	Devel	Eval
# sentences	22,962	500	500
# running words Chinese	165,999	3,522	6,085
# running words English	218,829	62,517	54,22
vocab. size Chinese	8,786	948	1,328
vocab. size English	7,944	3,878	2,347

speech. The domain of the data collection is limited to travel and booking information, i.e. the data collection contains speech about how to get to Cologne by train but no information about sports. The IWSLT 2005 corpus shown in Table 2 is a bilingual translation data collection featuring parallel sentences in Chinese and English from the travel and booking domain. Both data collections have in common that they focus on the single domain of travel and booking. The focus on a single domain is preferable because current state-of-the-art speech recognition and translation systems use domain specific models. For new sign language data collections, we propose to consider domains in which classifier signs are less likely to occur. Although sign language recognition systems will overcome the problem of recognizing classifier signs over time, classifier signs will remain difficult to automatically recognize and translate over an extended period of time.

Besides the choice of the domain, the average type-token-ratio is a key technical aspect that should be considered when creating new sign language data collections. The Verbmobil II corpus shows an average type-token-ratio of 69.01, and the IWSLT 2005 corpus an average type-token-ratio 18.8 respectively 27.54 for Chinese respectively English. The high average type-token-ratio of the Verbmobil II corpus is special to this corpus and not normally found in data collections used in speech recognition. Although the higher the type-token-ratio the better for a data-driven system, a type-token-ratio of 69.01 will not be achievable for sign language data collections in a reasonable time frame. Other well-known standard data collections in speech recognition such as the Wall Street Journal data collections (Paul and Baker, 1992) typically have type-token-ratios between 15 and 40. Taking into account the needs of data-driven speech recognition and translation, one goal in the recording of new sign language data collections should be an average type-token ratio of about 20.

The average type-token-ratio as such is a misleading figure, because the average can be biased by a small number

of very frequent tokens while the majority of tokens occurs only once or twice in a data collections. Therefore, the number of signs that occur only once in the data collection should be low. These singletons are in most cases named entities such as sign names or city names. In the Verbmobil II corpus and IWSLT 2005 data collections and several other benchmark databases for translation and speech recognition the percentage of singletons in the vocabulary is below 40%. This figure carries over to sign languages.

As already mentioned, the size of sign language data collections in terms of running signs or vocabulary size will not approach even the numbers given in Tables 1 or 2 over the next years. In order to keep the costs and time effort to create a sign language data collections that is also usable for data-driven approaches reasonable, we propose to aim for a vocabulary size that does not exceed 4,000 glosses (i.e. half the vocabulary size of IWSLT 2005 Chinese). Taking into account a desired average type-token-ratio of about 20 the envisioned data collections contains at most 80,000 running signs or 10% of Verbmobil II.

Data-driven translation systems typically exploit context information of words or complete phrases when translating a text from one language into another. The context used is typically limited to one sentence in order to limit computational cost. Therefore, data-driven translation translates one sentence of the source language e.g. sign language to an adequate sentence in the target language e.g. spoken language. This scheme requires bilingual sentence annotation as used in the IWSLT 2005 data collection. Unfortunately, the calculation of grammar inferred re-orderings of words is a computational expensive problem. Therefore, all used translation data collections limit the average sentence length to a range of 5 to 15 words in the source language. For a sign language data collection suitable for data-driven translation systems a similar bound should be used.

6. Conclusion

Most sign language data collection currently available for scientific research are of limited use to data-driven approaches to recognition and translation. We discussed the status of several sign language data collections available for scientific research from the point of view of data-driven speech recognition and translation. Based on the needs of data-driven approaches, we propose best practices for gloss annotation that ensure machine readable and adequate annotation of sign language while still allowing linguistically accurate annotation. Furthermore, we provide hard numbers for several technical aspects of data collections stemming from standard benchmark data collection of spoken languages. These hard numbers can act as references in the planning step for the creation of new sign language data collections.

7. Acknowledgments

This work has been partly funded by the European Community’s Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics - STREP) under grant agreement n° 231424 .

8. References

- J. Bungeroth, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, and L. van Zijl. 2008. The ATIS Sign Language Corpus. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.
- O. Crasborn and I. Zwitterlood. 2008. The Corpus NGT: An Online Corpus for Professionals and Laymen. In Crasborn, Hanke, Efthimiou, Zwitterlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages at LREC 2008*, pages 44–49, Paris. ELDA.
- O. Crasborn, E. van der Kooij, A. Nonhebel, and W. Emmerik. 2004. *ECHO Data Set for Sign Language of the Netherlands (NGT)*. Department of Linguistics, Radboud University Nijmegen.
- P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.
- T. Hanke and J. Storz. 2008. iLex — A database tool integrating sign language corpus linguistics and sign language lexicography. In *3rd Workshop on the Representation and Processing of Sign Languages at International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.
- C. Neidle, S. Sclaroff, and V. Athitsos. 2001. SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers*, 3(33):311–320.
- D. B. Paul and J. M. Baker. 1992. The Design of the Wall Street Journal-based CSR Corpus. In *DARPA SLS Workshop*, USA, February.
- 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*, volume 5 of *International Studies on Sign Language and Communication of the Deaf*. Signum, Hamburg, Germany.
- D. Stein, J. Forster, U. Zelle, P. Dreuw, and H. Ney. 2010. RWTH-Phoenix: Analysis of the German Sign Language Corpus. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta, May.
- U. von Agriss and K.-F. Kraiss. 2008. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. In *8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, September.

Building a corpus for Italian Sign Language: Methodological issues and some preliminary results

Carlo Geraci¹, Robert Bayley², Chiara Branchini¹, Anna Cardinaletti³, Carlo Cecchetto¹,
Caterina Donati⁴, Serena Giudice¹, Emiliano Mereghetti¹, Fabio Poletti³, Mirko Santoro³,
Sandro Zucchi⁵

¹University of Milan-Bicocca, ²University of California-Davis,
³University Ca' Foscari at Venice, ⁴Sapienza University of Rome, ⁵University of Milan.

Address: Carlo Geraci,
Department of Psychology
Università di Milano-Bicocca
Piazza Dell'Ateneo Nuovo, 1
20126 – Milano, Italy
E-mail: carlo.geraci@unimib.it

Abstract

The aim of this paper is to discuss some methodological issues that emerged during the creation of a corpus of data for Italian Sign Language, LIS. Data were collected from 10 cities spread across the country. 18 signers from each city have been recruited. They are native speakers of LIS or later-exposed to LIS and are divided into 3 age groups (19-38, 39-58, 59-78) of 6 signers each (3 males and 3 females). The methodology of data collection and transcription is similar to that used in previous studies of variation in American Sign Language (Lucas, Bayley & Valli 2001) and Australian Sign Language (Johnston & Schembri 2006), with some differences that we discuss. The corpus consists of various kinds of texts collected with different strategies: free conversation (45 minutes), elicited dialogues (about 5-10 minutes), narration (10 minutes) and a picture-naming task (42 items). For the transcription we adopted the ELAN software (Johnston & Crasborn 2006). Finally, a brief report on some preliminary results is presented.

1. Introduction

Since the earliest studies (Volterra, 1987), it clearly emerged that Italian Sign Language (LIS) has an impressive degree of variation. A few studies on lexical variation pointed out some phonological processes related to historical changes (Radutzky 2009, Geraci & Toffali, 2008), and a good number of geographical variants are reported in the most important LIS dictionaries (Radutzky, 1992 and DIZLIS, www.dizlis.it), while Bertone (2007) illustrates some register variations in the use of pronominal forms. However, systematic studies of this variation at various linguistic levels have not been carried out yet. The aim of this paper is to discuss some of the methodological issues that emerged during the creation of a corpus for LIS. Data collection is close to completion at the time of writing. A large-scale corpus has been constructed as part of a national research project on sociolinguistic variation in LIS (PRIN-2007). The core part of the project involves three universities: Sapienza University of Rome, University of Milan-Bicocca and University Ca' Foscari at Venice. As part of the project, the following studies are conducted (see also section 3): variation in the distribution of wh-signs, variation in the use of the 1/G handshape, variation in sign-order, lexical variation, variation in the use of the sign DEAF.

2. Issues in data collection

A first important issue concerns the selection of the cities where data were collected. On the one hand, our choice reflected the distribution of the urban population across the country; on the other hand, it reflected other

aspects of the culture and the language of the Italian Deaf community (for instance the presence in the past of important residential Deaf schools). Ten cities were selected, equally distributed across the country: four from the north (Bologna, Brescia, Milan and Turin), two from the centre (Florence, Rome), two from the south (Bari, Salerno) and two from major islands (Ragusa in Sicily, while data collection in Sardinia is imminent). The presence of two cities that are geographically close, namely Brescia and Milano, requires explanation. Despite their proximity, people from the two Deaf communities report clear differences in the use of LIS, possibly related to the existence in the past of an important residential school in Brescia.

For each city, we recruited a local contact person (usually with an active role in the deaf club) who was responsible for participant selection. A total of 180 signers from three age groups (18-30, 31-54, over 55) took part in the data collection. Both the local contacts and the participants were paid for taking part in the project, and participants also agreed to being recorded. For each city, data collection was completed in one day and a half (half a day for each age group).

The age grouping reflects the specific situation of Deaf education in Italy. Indeed, in 1977, a law of the Italian parliament stated that Deaf children could have access to mainstream education in ordinary schools. This law enabled parents to choose their children's education. Many parents (especially hearing parents) sent their deaf children to ordinary non-residential schools. Enrollment in non-residential schools undermined the only natural access to sign language for these children, and in a few years, almost all residential schools and special schools

for Deaf children closed. Hence, the older group (over 55) includes signers who attended residential Deaf schools, the middle group (31-54) includes signers who were at school age during the transition period, and the younger group (18-30) includes signers who had access to mainstream education. The protocol of data collection follows the main lines of those used for the creation of other SL corpora, in particular, the American Sign Language (Lucas, Bayley, & Valli, 2001) and Australian Sign Language (Johnston & Schembri, 2006) corpora. Data collection began with a 45-minute session of free conversation among three signers from the same age group. Then a session of question and answer elicitation followed, performed by pairs belonging to the same age group. The third task was an individual narration lasting approximately 10 minutes. Finally, each signer carried out a picture-naming task of 42 items. In contrast to Lucas, Bayley, & Valli, (2001), we opted for a smaller number of participants for the free conversation task, and we used three video cameras to record the session, one for each signer. One innovation of our study was a semi-structured question and answer task specifically designed to elicit wh-questions, a syntactic construction where variation was expected to occur (see section 3.1 and section 4). We introduced this session because it is unlikely that a number of wh-signs sufficient for a quantitative analysis would show up in free conversation signing. All participants performed the task in pairs: a scene was presented on a picture to one member of the pair. The other member could not see the picture but had to fill a form and recover the information needed by asking the partner. To illustrate, figure 1 depicts a car accident scene, while figure 2 shows the form to be filled out, which is very similar to the one Italian drivers fill out in case of small car accidents. By selecting a type of material that is mostly visual and a form that is familiar to signers, we strove to maintain as natural a situation as possible, even during a semi-structured elicitation procedure.



Figure 1: Car accident scene

In the individual narration session, signers were asked to tell some stories about their lives for about 10 minutes. In order to avoid the unpleasant feeling of signing right in front of a camera, and to reduce to a minimum the

potential effects of recording, the local contact was asked to play the addressee in this part of the data collection.

Figure 2: Insurance form

Finally, for the picture-naming task, 42 items from different lexical fields were selected in order to investigate variation in the lexicon of LIS. The list of the lexical fields includes: classifiers, compounds, color names, family names, fingerspelled words, initialized forms, month names, some specific signs known to be eligible for diachronic variation and new formations. Signers were shown an illustrated cardboard for each of the 42 items (see an example in figure 3) in a random order and were asked to name the represented object. During data collection no hearing researcher was present. One Deaf member of the research team was present at the very beginning of the free conversation session but he left the room when the exchange took off.



Figure 3: Picture-naming cardboard

3. Issues in data coding

Depending on the linguistic variable and on the part of the corpus under analysis (free conversation, elicitation session, etc.), different procedures have been adopted to investigate sociolinguistic variation in LIS. We report here those adopted in the study of the distribution of wh-signs and in the study of the I/G handshape variation.

For both studies, two Deaf native signers of LIS (each working on data from a different city) searched the tokens and did the first annotation of the variable by using the ELAN software (Johnston & Crasborn, 2006).

3.1 Distribution of wh-signs

Cecchetto, Geraci, and Zucchi (2009) conducted a qualitative in-depth study on wh-question formation in LIS and argued that wh-signs mostly appear in clause final position. To a lesser extent, wh-signs are reported to appear either in their argumental position, or reduplicated in situ and in clause final position. The aim of the study of the distribution of wh-signs is precisely to point out which factors are relevant in determining this variation. We analyzed the part of the corpus specifically designed to elicit questions. The first step in the annotation has been the identification of the utterances¹. In the first tier of the ELAN file, the coders simply had to delimit the utterances for that part of the corpus. This procedure has a double function: first, it facilitates the access to the database for further studies, and second it gives a rough measure of the productivity for each signer. The second step was to identify the utterances in which a wh-sign occurred and annotate the signs included in that utterance. The third step was to annotate the signs included in the utterance preceding the one containing the wh-sign, and the answer (if present) provided by the other signer (figure 4 illustrates the timetable of an annotation file). At this level, annotations were done in Italian and every wh-sign was specifically tagged with the ID “wh-” (e.g. “what” = wh-COSA, “who” = wh-CHI). This tag allows an easy identification of wh-signs via a simple search in the ELAN files. Although not immediately relevant for this phase of the study, further tags have been added in order to keep track of lexical variants for the wh-signs. In particular, a progressive number indicates alternative variants (e.g. wh-COSA, wh-COSA1, etc.), and a “0” right after the wh-tag indicates that the wh-sign is not the appropriate one (e.g. wh-0COSA means that the wh-sign for “what” is used instead of another wh-sign which is supposed to be more appropriate in that environment). These three steps were carried out by two Deaf native signers of LIS. In the fourth step, carried out by a CODA member of the research group, all the information coded with ELAN has been extracted in a worksheet file and further coding has been done. In particular, for each token, both linguistic and non-linguistic information has been added. As for linguistic information, we coded for the position of the wh-sign in the clause (reduplicated, before or after the predicate), utterance type (direct question, indirect question, echo question, alternative question, non-interrogative clause, pseudocleft), grammatical

¹ We are aware that the definition of utterance is controversial both for sign and spoken languages, and that native users of a language have different intuitions about where an utterance ends (see Barrett, 2008 for a recent discussion of this issue in spoken languages).

function of the wh-sign (subject, object, adjunct, etc), wh-type (who, what, when, etc.). As for social information, we coded for geographical origin, gender, presence of Deaf people in the family (parents, relatives or none), education (kindergarten, primary school, middle school or higher education) and work experience (blue collar, white collar, professional or student).



Figure 4: View of the ELAN workspace

3.2 1/G handshape variation

Our aim in the study of phonological handshape variation is to replicate a similar study conducted on ASL by Lucas, Bayley, & Valli (2001). The two crucial methodological differences between our study and that of Lucas et al. are the use of a dedicated camera for each signer instead of a single camera for all the signers involved in the conversation, and the use of ELAN for the coding. Differently from the study of the distribution of wh-signs, where the coding was done in two separate steps, in this case all the coding is done within the ELAN file. This has been made possible by using multiple tiers organized hierarchically (see figure 5). The organization in figure 5 may look complicated but, coding was in fact quite simple since most of the tiers adopt a controlled vocabulary, resulting in a pull-down menu. This choice allows the coder to control for the effects both of single features (such as number of selected fingers, or their hooked vs. straight status) and of combinations of features (i.e. groups of handshapes). In figure 5, the first two tiers, namely the main tier (fo1, i.e. Firenze Old signer number 1) and the GLOSS tier are devoted to highlight the sign with the 1/G handshape, the preceding sign and the following sign. The rest of the relevant tiers depends on the GLOSS tier and can be grouped in three main sets, 1-Dhand, 1-Ante Pause, 1-Post Pause, which provide information about the dominant hand, the preceding and following sign, respectively. The main characteristic of these tiers is that each of them is made up with a controlled vocabulary. For sake of exposition we illustrate here the case of the set of 1-Dhand tiers. The 1-Dhand tier specifies the number of selected fingers (other than the index finger and thumb) for the variable token (0, 1, 2, 3). The 1-Dindex tier specifies whether the index is extended, closed (as in the S handshape) or hooked. The 1-Dthumb specifies whether the thumb is extended or not, while the 1-Dhooked specifies whether the selected fingers are extended or hooked. Finally, the 1-Class tier specifies

the grammatical class of the token (pronoun, noun, verb, adjective, adverb, functional sign). The advantage of this coding is immediate once the data are extracted for statistical analyses. Indeed, each tier is converted into a factor group already in columns.

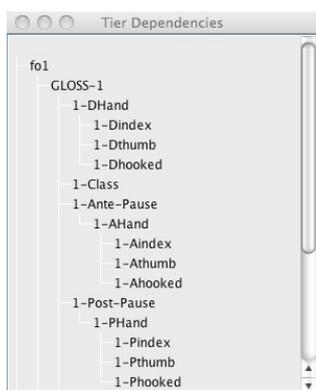


Figure 5: 1/G handshape study tier dependencies

Furthermore, each factor group (including the dependent variable) is already fully specified, since its values come from the close array determined by the controlled vocabulary.

4. Preliminary results: the case of wh-signs

Although the coding for the cities has not yet been completed, some preliminary results about the distribution of wh-signs in LIS are worth mentioning. In particular, the data reported in table 1 are from three cities (Bari, Bologna, and Turin), and illustrate the percentages of the distribution of wh-signs occurring reduplicated (in situ and in clause final position), before and after the predicate.

The general observation made by Cecchetto, Geraci and Zucchi (2009) that the most natural position for wh-signs is the right periphery of the clause is confirmed for all age groups. Furthermore, the data nicely show a diachronic pattern of development in that the proportion of wh-signs occurring in preverbal position decreases across the three age groups from 35% to 17% and then further to 10%. This reduction is compensated by a neat increment in the postverbal positioning of wh-signs and in a moderate increment of reduplicated forms.

Age	After	Before	Reduplicated
Old (over 55)	49%	35%	16%
Middle (31-54)	63%	17%	20%
Young (18-30)	68%	10%	22%

Table 1: Distribution of wh-signs by age groups

5. Conclusions

In this paper, we addressed some of the major issues related to the collection of a corpus for LIS and one

preliminary result emerging from the analysis of such corpus. Although the basic structure of our project is similar to that used in other projects that have collected sign language corpora, we introduced some innovations such as the use of a camera to record each individual signer's production, more structured elicitation sessions to elicit particular syntactic constructions and specific coding steps motivated by the use of the ELAN software as a main tool for data coding.

6. References

- Barrett, R. (2008). Linguistic differentiation and Mayan language revitalization in Guatemala. *Journal of Sociolinguistics* 12(3), pp. 275--305.
- Bertone, L. (2007). *La struttura del sintagma determinante nella Lingua dei Segni Italiana (LIS)*. PhD. Dissertation, University Ca'Foscari at Venice.
- Cecchetto, C., Geraci, C., & Zucchi, S. (2009). Another way to mark syntactic dependencies: The case for right peripheral specifiers in sign languages. *Language*, 85(2), pp. 1--43.
- Geraci, C., Toffali, L. (2008). Tendenze conservatrici e innovative nell'uso delle lingue: la variabile dell'età nella Lingua dei Segni Italiana. In G. Bella, D. Diamantini (Eds.), *La qualità della vita nella società dell'informazione*. Milano: Guerini e associati, pp. 97--115.
- Johnston, T., Crasborn, O. (2006). The use of ELAN annotation software in the creation of signed language corpora. In *Proceedings of the EMELD '06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art*. Lansing, MI.
- Johnston, T., Schembri, A. (2006). Issues in the creation of a digital archive of a signed language. In L. Barwick, & N. Thieberger (Eds.), *Sustainable data from digital fieldwork*. Sydney: University of Sydney Press, pp. 7--16.
- Lucas, C., Bayley, R., & Valli, C. (2001). *Sociolinguistic variation in American Sign Language*. Washington, D.C.: Gallaudet University Press.
- Radutzky, E. (Ed.). (1992). *Dizionario bilingue elementare della Lingua dei Segni Italiana LIS*. Roma: Edizioni Kappa.
- Radutzky, E. (2009). Il cambiamento fonologico storico della lingua dei segni italiana. In C. Bertone & A. Cardinaletti (Eds.) *Alcuni capitoli della grammatica della LIS. Atti dell'Incontro di studio "La grammatica della Lingua dei segni italiana"*. Venezia: Cafoscarina, pp. 17--42.
- Volterra, V. (Ed.). (1987). *La lingua dei segni italiana. La comunicazione visivo-gestuale dei sordi*. Bologna: Il Mulino.

Acknowledgements

The work reported in this paper has been funded by PRIN 2007 project "Dimensions of variation in Italian Sign Language". The Italian team warmly thanks Ceil Lucas and Adam Schembri for sharing their expertise in various phases of this project.

SiS-Builder: A Sign Synthesis Support Tool

Theodore Goulas, Stavroula-Evita Fotinea, Eleni Efthimiou, Michalis Pissaris

ILSP-Institute for Language and Speech Processing / ATHENA R.C.

Artemidos 6 & Epidavrou, 151 25 Maroussi, Athens, Greece

tgoulas@ilsp.gr, evita@ilsp.gr, eleni_e@ilsp.gr

Abstract

SiS-Builder is a web based tool, developed in the framework of the DICTA-SIGN project in order to cover for the need of creating sign language (SL) lexical resources, adequate for sign synthesis performed by a signing avatar (virtual signer). The tool's initial function was to automatically generate SiGML transcriptions of HamNoSys strings, as well as the relevant transcription files, by providing the HamNoSys characters of a sign. SiS-Builder provides an environment, which is accessible by everyone and allows interaction without requiring any special installations on the client side. The tool enables users to create or review HamNoSys notations and SiGML scripts of sign lemmas on line, switch between SiGML data and HamNoSys notations by selecting the wished function, review already created lemmas and proofread avatar performance, also viewing the video of the performed sign. It makes then possible to experiment with the results of synthesis of lexicon items, by either consulting the HamNoSys sequence, for those familiar with the HamNoSys syntax, or animating the results through the avatar with the use of the SiGML script.

1. Introduction

SiS-Builder is an online tool initially developed to serve needs of the DICTA-SIGN project, in relation to creation of SL lexical resources and research work on synthesis and animation. The most prominent need that led to its design and implementation was the requirement to generate SiGML transcriptions of HamNoSys strings in order to feed the sign synthesis avatar of the University of East Anglia (UEA) (<http://vh.cmp.uea.ac.uk>) (Elliot et al., 2000; Elliot et al., 2004a). In the course of its implementation, SiS-Builder was enriched with a number of functionalities that provide a complete environment for creating, editing, maintaining and testing lexical resources of sign languages, appropriately annotated for sign synthesis and animation. In the rest of the paper the components and functionalities of the tool will be separately presented. The tool is based on open source internet technologies to allow for easy access and platform compatibility, mostly exploiting “php” and “java script”, and is accessible through the following URL: <http://speech.ilsp.gr/sl/>.

2. A Sign Synthesis support tool

Sign synthesis and animation have been accused in the past for lacking the naturalness of human signing, equally due to avatar motion performance and complete absence of the non-manual articulation elements from synthetic signing (Karpouzis et al., 2007). Research on sign synthesis is currently experimenting with ways to overcome these weaknesses by improving performance of manual articulation and also by implementing non-manual features (Elliot et al., 2004b). However, the demand for properly coded lexical data to support sign synthesis is increasing, the latter being a time consuming procedure, performed only by human coders (Elliot et al., 2008; Fotinea et al., 2008). SiS-Builder is a tool developed to facilitate creation of lexical resources for sign synthesis, enabling multiple users to create and test their own data sets. As such, the tool is Internet based, free to be accessed by anyone, with no special installation requirements on the client side. It provides a GUI, via which users can automatically create SiGML scripts to be used by the UEA avatar animation engine,

either by entering HamNoSys strings (Prillwitz et al., 1989; Hanke, 2004) of signs already stored in a properly coded lexical database, or by creating HamNoSys annotated lemmas online, using the relevant SiS-Builder function (Figure 1). Once the HamNoSys coded data are provided, automatic conversion of the characters and creation of the corresponding SiGML script takes place. The so created script can be stored upon demand on the SiS-Builder server and be ready for future use.

Editing is also possible on already stored SiGML scripts, which allows for immediate presentation of the modified resource by the avatar. In order to store the new lexical item, however, it is obligatory to provide the relevant HamNoSys descriptions which in turn will be converted and stored as a SiGML script.

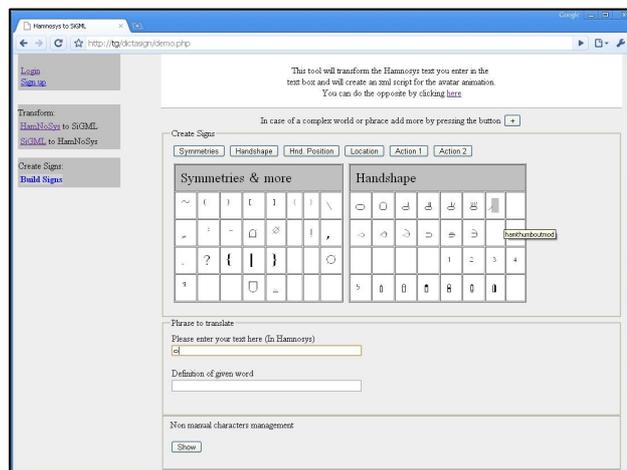


Figure 1: Creating HamNoSys notations online

3. Converting HamNoSys to SiGML

Currently, users may animate a simple sign or a sign phrase consisting of up to four lexical items. To do this, the user has to enter the relevant HamNoSys notations in the field labeled as “Please, enter you text here” (Figure 2) in the GUI. After having entered the HamNoSys annotated string in the proper field, the user may add the non manual characters of the sign she/he is dealing with,

by first selecting from the “*Non manual characters*” section the “*check*” buttons she/he needs, and then choosing the “*show*” button in the “*Non Manual Characters Management*”. The user may choose the non-manual features she/he needs to describe a sign, from an (almost) exhaustive list of non-manual characters, been implemented by UEA. To achieve a performance as close to natural as possible, the user may choose a variety of features, such as combined movements for the head like tilt left and swing right. The user may apply the same procedure for more than one sign until she/he proceeds to the final step, which is creation of the relevant SiGML script. An indicative part of the implemented non-manual features, is shown in Figure 3.

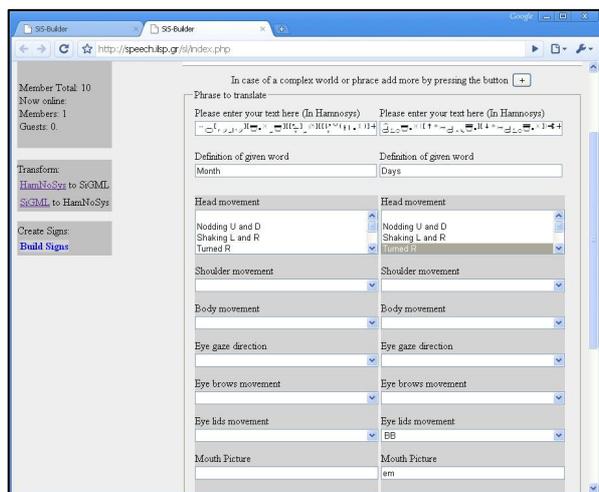


Figure 2: Converting HamNoSys to SiGML

```

signlnonmanual.dtd
<!ENTITY % nose_abbrev
  "WR | TW | WI"
>
<!ENTITY % nose
  "%nose_abbrev;
  WR_wrinkled_nose |
  TW_twitching_nose |
  WI_widened_nostrils
  "
>
<!ENTITY % mouth_gesture_abbrev
  " D01 | D02 | D03 | D04 | D05 | D06 | D07 | D08 | D09
  | J01 | J02 | J03 | J04
  | L01 | L02 | L03 | L04 | L05 | L06 | L07 | L08 | L09 | L10
  | L11 | L12 | L13 | L14 | L15 | L16 | L17 | L18 | L19 | L20
  | L21 | L22 | L23 | L24 | L25 | L26 | L27 | L28 | L29 | L30
  | L31 | L32 | L33 | L34 | L35
  | C01 | C02 | C03 | C04 | C05 | C06 | C07 | C08 | C09 | C10
  | C11 | C12 | C13
  | T01 | T02 | T03 | T04 | T05 | T06 | T07 | T08 | T09 | T10
  | T11 | T12 | T13 | T14 | T15 | T16 | T17"
  >
  
```

Figure 3: Part of the implemented non-manual entities

4. Converting SiGML back to HamNoSys

If wished, SiS-Builder provides for the option to convert a SiGML script back to the corresponding HamNoSys notation. The results of this transformation are depicted

in Figure 4. As already mentioned, HamNoSys strings are visualised either by converting data in SiGML format or by selecting a validated lexical item from a list.

5. Repository of sign resources

Registered users of the SiS-Builder environment have access to the signing resources (signs, concepts and phrases) repository incorporated in the environment.

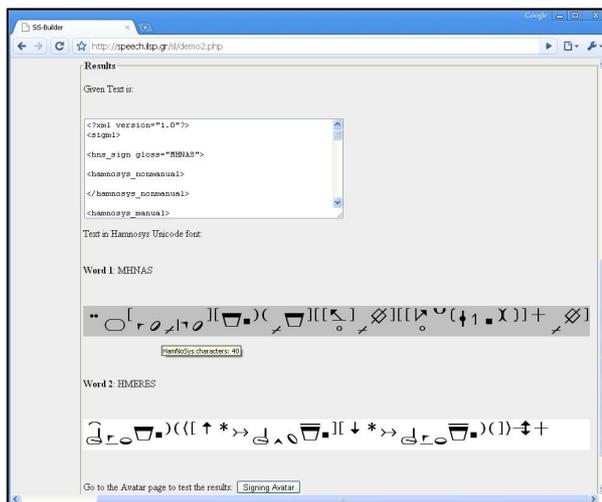


Figure 4: Converting SiGML back to HamNoSys

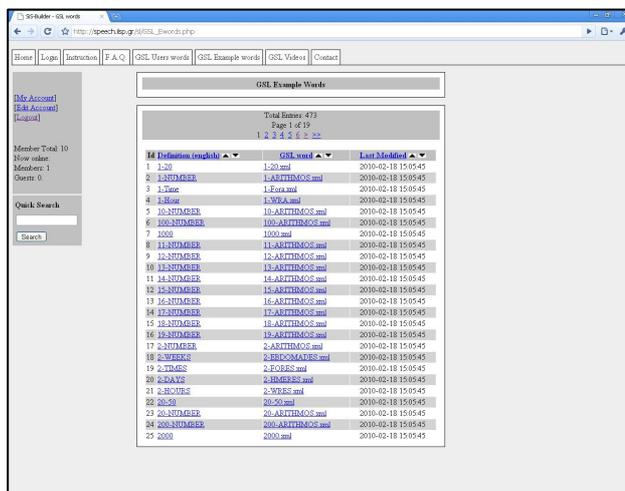


Figure 5: Repository of signs

Users can go through the available list of signs or search for the lexical resource they are interested in, by exploiting the “*Quick-Search*” function, visualised on the left hand side menu of the screen. Figure 5 depicts an instantiation of the GSL repository of lexical resources.

5.1 Users’ repository

Registered users of the SiS-Builder environment, are provided with their own repository space. This facility allows users to store, experiment with and modify the lexical resources they have previously created, until they achieve satisfactory descriptions for synthesis, corresponding to appropriate animation performance.

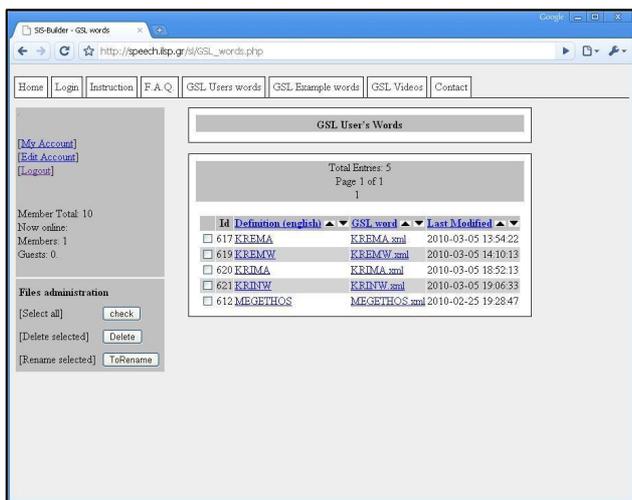


Figure 6: User’s repository

Figure 6 depicts a specific user’s repository, where modifications of lexical resources may be performed.

5.2 Video repository

In order to facilitate users to create lexical resources for avatar animation or to make it possible to compare natural signing resources, with the avatar’s SL performance, a video repository is offered to registered users, containing authentic signing data.

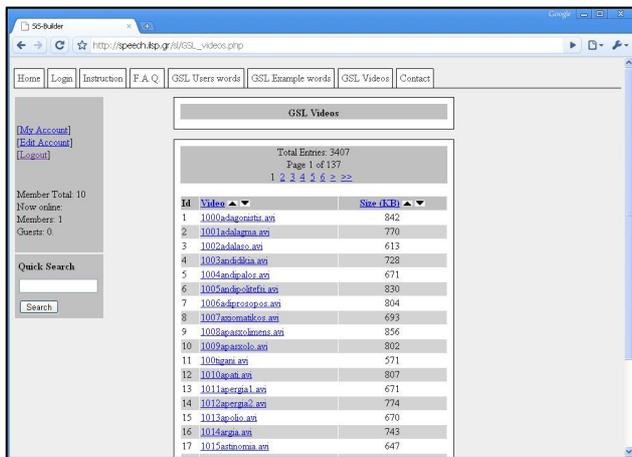


Figure 7: GSL videos repository

To demonstrate the comparison utility of the specific functionality, Figure 8 depicts an instantiation of video presentation of a specific lexical item (the GSL sign for “wild” in this example), while Figure 9 instantiates the same lexical item when visualised by the UEA avatar. Search in the video repository is possible either by means of the English form of a lexical item, its Greek form or the name of the corresponding video file. A partial search query is also possible. Users can search for example the item “lawyer” by typing either “lawyer” or “law” or the name of the corresponding video file, if this is known to them, i.e. typing “1061law.avi” or “δικ” or “δικηγόρος”, if they make the same query in Greek. Search results are then presented to the user in a similar way, as the list shown in Figure 8.

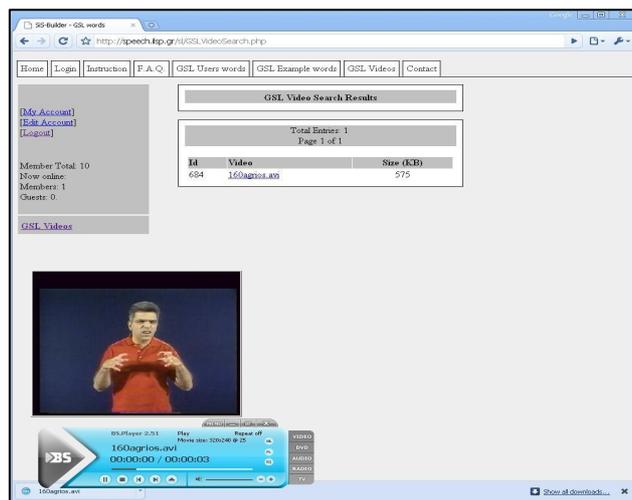


Figure 8: GSL video for the sign “wild”

Lexical resources created in the SiS-Builder environment may feed the UEA signing avatar accessible at: <http://vhg.cmp.uea.ac.uk/tech/jas/095z/SPA-framed-gui-win.html>. Examples of the signing avatar female model, Anna, are presented in figures 9, 10 and 11 below. In Figure 9 one can observe the avatar environment, which is designed and developed by UEA. On the left hand side of the screen, the SiGML script created by SiS-Builder, describes the sign for “wild”, where avatar performance is directly comparable to natural representation of the same sign in the video of Figure 8.

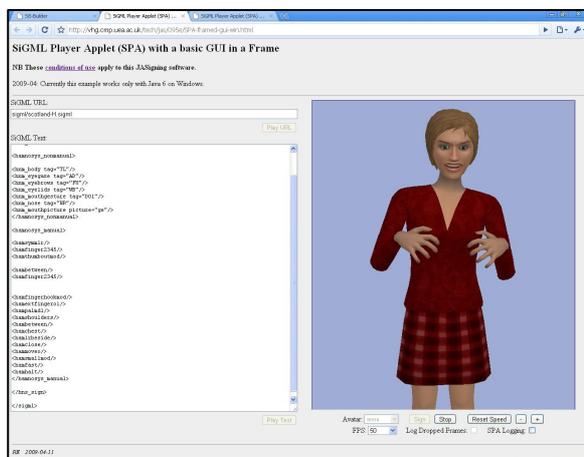


Figure 9: Avatar signing the GSL sign “wild”



Figure 10: Avatar signing (zoom) the GSL sign “cat”, demonstrating non manual features implementation

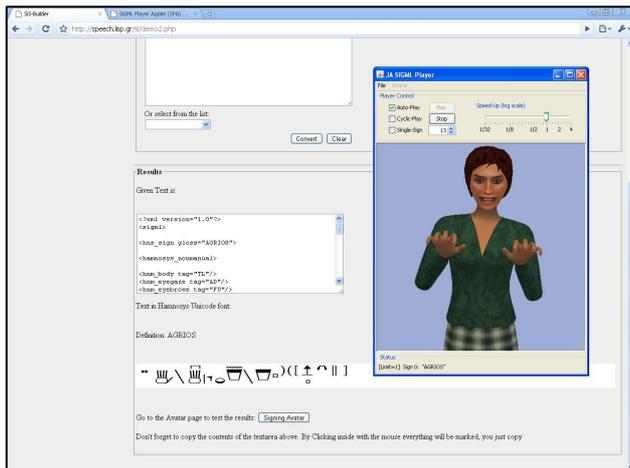


Figure 11: On the background SiS-Builder's results of the GSL lexical resource "wild", while a synonym of the sign represented in Figures 8 and 9 is performed by the UEA avatar

6. Conclusion

SiS-Builder was developed to assist creation of lexical resources appropriately coded for sign synthesis and animation (Efthimiou et al., 2006) in the framework of a specific research task. In this respect, SiS-Builder makes use of the HamNoSys notation system and the SiGML scripting language to speed up lexical resources creation and cover needs for multilingual synthesis. Implementers enriched the initial environment, though, with a number of functionalities which allow for the long term use of the tool by a wide range of users in the scope of creating lexical resources of SLs, fully coded for sign phonology (Kouremenos et al., to appear).

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

8. References

- Efthimiou, E., Fotinea, S-E., and Sapountzaki, G. (2006). Processing linguistic data for GSL structure representation. In *Proceedings of the Workshop on the Representation and Processing of Sign Languages: Lexicographic matters and didactic scenarios*, Satellite Workshop to LREC-2006 Conference, May 28, pp. 49--54.
- Elliott, R., Glauert, J.R.W., Kennaway, J.R., and Marshall, I. (2000). Development of Language Processing Support for the Visicast Project. In *ASSETS 2000 4th International ACM SIGCAPH Conference on Assistive Technologies*, Washington DC, USA, 2000.
- Elliott, R., Glauert, J.R.W., Jennings, V., & Kennaway, J.R. (2004). An Overview of the SiGML Notation and SiGMLSigning Software System. In O. Streiter and C. Vettori (Eds.), *Proceedings of 1st Workshop on Representing and Processing of Sign Languages*, LREC 2004, Lisbon, Portugal, pp. 98--104.
- Elliott, R., Glauert, J.R.W., & Kennaway, J.R. (2004). A Framework for Non-Manual Gestures in a Synthetic Signing system. In Keates, S., Clarkson, P.J., Langdon, P., & Robinson, P., (Eds.) *Proceedings of 2nd Cambridge Workshop on Universal Access and Assistive Technology (CWUAAT)*, Cambridge, UK, 2004, pp. 127--136.
- Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I. and Sáfár, E. (2008). Linguistic Modelling and Language-Processing Technologies for Avatar-based Sign Language Presentation. In Efthimiou, Fotinea, Glauert (eds) *Emerging Technologies for Deaf Accessibility in the Information Society*, Special Issue, *Journal of Universal Access in the Information Society*, Vol 6, No 4, pp. 375--391.
- Hanke, T. (2004). HamNoSys - representing sign language data in language resources and language processing contexts. In *LREC 2004, Workshop proceedings: Representation and processing of sign languages*. Paris: ELRA, 2004, pp. 1--6.
- Fotinea, S-E., Efthimiou, E., Karpouzis, K., Caridakis, G. (2008). A Knowledge-based Sign Synthesis Architecture. In Efthimiou, Fotinea, Glauert (eds) *Emerging Technologies for Deaf Accessibility in the Information Society*, Special Issue, *Journal of Universal Access in the Information Society*, Vol 6, No 4, pp. 405--418.
- Karpouzis, K., Caridakis, G., Fotinea, S-E., and Efthimiou, E. (2007). Educational Resources and Implementation of a Greek Sign Language Synthesis Architecture. *Computers and Education*, Elsevier, Volume 49, Issue 1, August 2007, pp. 54--74, electronically available since Sept 05.
- Kouremenos, D., Fotinea, S-E., Efthimiou, E., and Ntalianis, K. (2010). A Prototype Greek Text to Greek Sign Language (GSL) Conversion System. *Behaviour & Information Technology Journal (TBIT)*, (in print).
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T. and Henning, J. (1989), HamNoSys. Version 2.0. Hamburg Notation System for Sign Language: An Introductory Guide, Hamburg: Signum Verlag, 1989.

DGS Corpus & Dicta-Sign: The Hamburg Studio Setup

Thomas Hanke, Lutz König, Sven Wagner, Silke Matthes

Institute of German Sign Language and Communication of the Deaf, University of Hamburg
 {thomas.hanke,lutz.koenig,sven.wagner,silke.matthes}@sign-lang.uni-hamburg.de

Abstract

We describe the setup for a mobile studio used for data collection of both the DGS Corpus project and the DGS part of the Dicta-Sign corpus. This includes camera positioning and the software used to conduct a recording session as well as the rationale behind the decisions taken in that respect, but we also investigate the challenges the moderator of a session has to face.

1. Introduction

Not taking into account budget restrictions, the setup of a sign language studio always is a balancing act between high quality recordings on the one hand not to make the transcription process even more complicated than it is anyway and possibly to enable automatic processing of the recordings, and on the other hand an environment where the informants still feel comfortable enough so that the recording situation does not have too much impact on the signing. In the case of the DGS Corpus project, an additional constraint is that the studio is to be relocated twelve times over the course of two years as it was decided to make the recordings in the regions instead of inviting participants to one central place to avoid dialectal mixing. One of the implications of this approach is that the studio is operated by non-specialist deaf fieldworkers with limited time available for training.

2. Overall Setup

The elicitation setting for both DGS Corpus and Dicta-Sign involves two informants interacting in different ways with each other and a moderator (the fieldworker from the region) leading the session.

During a recording session the pair of informants is sitting facing each other at an approximate distance of three meters. Camera positions in the studio guarantee that each informant is being filmed separately while there exists another camera to film the overall scene. Elicitation material to be shown to the informants is presented on screens in the middle between the signers



very close to the ground in order not to interfere with their views of each other. The moderator is sitting next to the informants as shown in the picture below. He/she introduces the tasks and observes the conversation, but only interferes with the conversation if absolutely necessary. Monitors in front of the moderator display the “Session Director” (described below) as well as what can currently be seen on the informants’ screens.

3. Camera Positions

The camera setup we finally ended up with consists of seven cameras altogether, three on each informant and one for the whole scene including the moderator. Two HD cameras provide frontal views of the informants while birds-eye cameras capture each informant’s signing from above to help human transcribers to interpret the signing. Additionally two stereo cameras mounted on top of the frontal-view cameras capture the signing in parallel. They are to provide footage that allow image analysis to reconstruct 3D information and help automatic processing. The seventh camera is again an HD camera which captures the whole scene, i.e. both informants as well as the moderator interacting with the informants to give the transcriber a quick overview and to help him/her to exactly identify interactions between the three participants.

One of the advantages of using HD cameras is that extra cameras for close-ups, e.g. on the signers’ faces, are not necessary as the spatial resolution of the those parts of an HD image still is comparable to what can be achieved with SD cameras – without the need to track positions.

In earlier projects the informants were also seated opposite each other but at a slight angle. The cameras were positioned at eye height on the side of each informant. The drawback of this approach is some informants constantly target their signing back and forth between the addressee and the camera, which makes it difficult to identify important aspects as body shifts in the discourse situation. Pre-tests were therefore conducted testing different seating arrangements and camera positions. It was found that a setting is optimal where the informants are directly facing each other in order to assure a relaxed conversational situation. With this, three different camera positions are possible: on the side or above the other informant’s head, or between the informants. The last option requires a large distance between the informants or shooting at wide angle. In

addition, to avoid blocking the informants' views, the cameras would need to be mounted quite low, having to shoot upwards, something our transcribers did not like at all. The side position provides a view on the informant at eye height but at a side angle, which makes body shift and eye gaze tracking difficult for the transcriber as well as for semi-automatic processing. For the Hamburg studio setup it was therefore decided to use a setting with the main cameras positioned above and behind the heads of the informants, which provides a view from the front but with a slight angle from above. The pre-tests revealed that with a distance of approximately three meters, the distortion introduced by the elevated position of the camera does not negatively affect the transcription from video. Instead, this setting provides a front view of the informant similar to the addressee's, allowing identification of body shifts as well as eye gaze direction more easily. At the same time, especially with the monitors being located on the floor between them, the informants did not feel this to be an unnatural distance for their interaction.

4. Procedure of Elicitation Sessions

4.1 The moderator's role and presentation of the stimuli

The fieldworker of a certain region also serves as the moderator during the elicitation sessions. In order to avoid influences on the language production it is crucial that the moderator is a Deaf person and that no other (hearing) person is present in the studio throughout the session.¹ The moderator is responsible for the presentation of the individual tasks as well as a smooth run of the whole session.

Each task is briefly introduced by the moderator, followed by a detailed explanation and instruction for the informants, presented in a DGS video clip on the informants' monitors. While these pre-recorded explanations ensure that all informants get exactly the same information and that nothing is left out, further clarifications given by the moderator might become necessary for some of the informants before starting a task. The materials used as stimuli during a task comprise of different media formats, including pictures, drawings and video clips. They are shown as slides on the informants' monitors in a semiautomatic presentation, partly one slide following the other at a fixed speed, partly controlled by the moderator (e.g. allowing for in-between questions by the informants). Depending on the individual task the presentation of this material might be identical or different for the two informants.

The aim of the tasks and the stimuli used is to evoke a conversation between the two informants, while the moderator should observe the conversation and only interfere if absolutely necessary. However, it is the moderator's duty to check the time used for each task in order to ensure that enough time is left for the rest of the sessions. For both Dicta-Sign and DGS corpus extra tasks were planned that can be included or left out

depending on the time left.

Leading the elicitation sessions and taking care of every aspect required leaves a heavy responsibility with the moderator. Training sessions are therefore required to fulfil this task. However, with limited time available prior to the elicitation and especially with long elicitation sessions to be performed (seven hours for the DGS corpus elicitation), a (semi)automatic control of the session should be implemented wherever possible. A custom software was therefore developed in order to support the moderator in his/her work and to ensure a smooth work flow.

4.2 Session Director

Session Director allows the moderator to present slides to the informants. The screen configuration for the presentation of slides is one screen facing each informant and for the moderator one screen to run Session Director and two additional screens to observe the presentations on the informants screens. This means that the moderator can also see what is shown to the informants.²

In addition to the list of tasks, Session Director's main window shows detail of the task current worked on. This includes a progress bar showing the time already spent on the task in relation to the time planned in as well as the sequence of subtasks, such as the instructions to be presented to the informants, stimuli presentations and conversations between the informants. While the sequence of events is predefined, it is the moderator clicking start buttons on the screen to activate the next step. This allows the moderator to check if all explanations have been understood. In case, s/he can decide to repeat instructions on screen or to rephrase the task himself/herself before moving on.

All the moderator's interactions with Session Director are logged with time stamps. This allows us to determine automatically where on the videos certain tasks (or pauses) can be found and also to conclude from our knowledge of the tasks who presumably is the active signer at a given point in time and e.g. to use this information to automatically zoom the displayed video onto that person.^{3,4}

It is neither doable nor desirable to keep each task to the planned duration. It is, however, necessary to keep the total session time close to the plan as the informants may need to catch their train back home at the end of the day etc. For this purpose, Session Director shows another window giving the actual time, the elapsed time in the session and the time before/behind plan. The window changes colours to signal when deviations exceed a

² Technically, the informants' screens are computers (iMac 24") with a second monitor mirroring the content to the moderator. They run a slide show presentation program (Apple Keynote) remote-controlled (via AppleEvents) by Session Director running on the moderator's iMac. The integration of Keynote allows us to use its full repertoire of features such as transition effects where appropriate.

³ Notes that the moderator takes during the session (lower right of the main window) are also output into that log file. While we did not expect this to be much used, one of the three moderators did use this possibility regularly.

⁴ The log also easily translates into tagging in our transcription environment iLex, allowing links from the transcript to the task description and vice versa.

¹ A Deaf technician monitors the recording equipment from next door. While the informants know that, they do not see him/her during the session.

certain threshold and require corrective action. If the session is well behind for some reason, it might no longer be possible or advisable to reduce the signing time for each task to a minimum, but to skip entire subtasks or even whole tasks. Session Director supports the moderator in these decisions with marking lower-ranked tasks that could be skipped.

In general, the order of tasks, including breaks, is predetermined by the session description. The moderator has, however, the freedom to rearrange tasks or to change the expected duration of the session Session Director measures progress against. This might for example become necessary if one informant arrives late. But we have also experienced the opposite: Informants had so much fun with the tasks that the session was well behind, but they preferred to stay an hour longer in order not to miss any of the tasks still to come. With the moderator redefining the session duration, Session Director went back from “condition red” to “condition green” so that the moderator could relaxedly monitor progress without being constantly reminded to rush or skip optional parts.

4.3 Session Description Files

When launched, Session Director loads an XML file describing the session. For each task and subtask, it defines the expected and maximum acceptable duration, the text of the user interface elements visible in Session Director and of course the ids of the slides to be shown to the informants, either as a common set or separately for each informant. Furthermore, it defines the relative importance of each task which Session Director will eventually use to mark tasks that can be skipped. In addition, text can be entered that will be displayed alongside with the task detail. This could be reminders to the moderator what questions to ask to get a discussion going should that turn out to be necessary for a specific task.

Separating the description from the program also makes it easy for us to produce session-specific files. We currently use this to arrange for alternating tasks: Some tasks are only used in every second or fourth session.⁵ The moderator can then use the file produced for a specific session to prepare for the session, e.g. by adding keywords to indicate the informants’ hobbies and such, information to be used in the warm-up phase or whenever some intervention by the moderator becomes necessary.

4.4 Training

While the Session Director user interface is quite straightforward to use, time management for the data collection sessions remains a demanding task and requires that the moderator is familiar with the program under all conditions. Moderators are introduced to the program within the fieldworkers training courses. Right away, they have to use the program to manage rehearsal sessions with a mixture of cooperative and not-so-easy “informants”. Time management remains the most complex aspects, and feedback from the moderators has led to some modifications as to when the colour-coded

⁵ The Keynote slides document is the same for all variations of the session: It contains the material for all variations, with some not being called in each session.

time reminders appear. Feedback from the first sessions also made us introduce a Pause function to halt the task time should a spontaneous break become necessary (while of course the session clock is not stopped).

The fieldworker’s manual also documents Session Director as well as the essentials of time management. In addition, the slides as well as training version of Session Director⁶ are made available to the moderators.

5. Technical Details

The HD cameras we use are 3 Sony EX3 and 2 Sony HRX-MC1P (for the birds-eye views), all recording at 1080i25. These cameras store data locally on memory cards (only used as a fallback solution) and at the same time stream into MacPros running FinalCutPro (via SDI connections for the EX3 and HDMI connections for the HRX-MC1P.)

The stereo cameras mounted on top of the EX3s are PointGrey Bumblebee 2 models capturing 640x480p48 for each channel. They are connected via FireWire to MacPros running capture software provided by Dicta-Sign partner University of Surrey (currently under Windows XP).⁷

With total data rates of 700 GBytes per hour, we are not able to transfer the data to the Hamburg server before the next session starts. Instead, we swap hard disks (Raids consisting of two 2TBytes hard disks for each computer) after each fourth session and transport them back and forth in special suitcases. (Local backup is available just in case.)

For the mobile studio in the described configuration, we need a room of at least 5m x 5m optimally with a ceiling height of at least 3m and a smaller room next door. However, the larger the room, the less packed it is, the easier it is to make the informants feel comfortable.

For relocating the studio, we have transport cases for all equipment including spare parts (including one HRX-MC1P, one MacPro, one iMac). Deinstallation, actual transportation and installation at the next site is organised by an external service provider.

Session Director is available free of charge for MacOS X only as it heavily relies on MacOS-specific functionality.⁸ Source code is available upon request. As most of the user interface of Session Director is determined by the session description XML files, no localisation is necessary. The manual currently is available only in German, but an English version should become available in the future.

⁶ The training version does not actually remote-control the slides computers as that would require a multi-computer setup at the moderator’s home.

⁷ Remote control for FinalCutPro and the Bumblebee recording software was implemented and integrated into Session Director but is currently not used. It turned out that too many things can happen that need immediate attention than could be handled by the moderator. We therefore decided to have a technician on-site which allows the moderator to concentrate on dealing with the informants and managing the session.

⁸ The training version, however, is also available for Windows.

Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135 and from the German Academies of Science Programme.

Example: Excerpt from a Session Description XML file and the corresponding view in Session Director:

```
<session name="DGS-Korpus Hamburg" standard_rank="5" green_threshold="-30" yellow_threshold="0" red_threshold="20" maximum_duration="420"
keynote_a="a.local" keynote_b="b.local" >
```

```
<recording id="start 1" name="Aufnahme ein" action="start" average_duration="1">
<checklist text="Alle Kameras nehmen auf"/>
<checklist text="Wichtig: Über den Monitoren einmal laut Klatschen (zur Filmsynchronisation)!"/>
</recording>
```

...

```
<task id="27" name="Ablaufbe. + Reisesgeschichte" average_duration="25" maximum_duration="30" rank="2" >
<presentation id="27.1" name="Aufgabenverteilung" source_a="314" source_b="316" duration="0:33" />
<subtask id="27.2" name="Ablaufbeschreibung" target="A" >
<presentation id="27.2.1" name="Aufgabenerklärung" source_a="318" source_b="332" duration="0:41" />
<presentation id="27.2.2" name="Themen (1:40 min)" source_a="320" source_b="430" duration="1:39" />
<presentation id="27.2.3" name="Themenübersicht" source_a="322" source_b="433" average_duration="2" maximum_duration="2" />
<narration id="27.2.4" name="Ablaufbeschreibung wird gebärdet" source="2" average_duration="7" maximum_duration="9" />
</subtask>
<subtask id="27.3" name="Reisesgeschichte (Vor dem Gebärdn auf schwarz schalten!)" target="B" >
<presentation id="27.3.1" name="Aufgabenerklärung" source_b="330" source_a="328" duration="0:33" />
<presentation id="27.3.2" name="Reisesgeschichte (2:30 min)" source_b="334" source_a="421" duration="2:26" />
<narration_with_stimulus id="27.3.3" name="1. Abschnitt der Reisesgeschichte / schwarz" source_b="353;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.4" name="2. Abschnitt der Reisesgeschichte / schwarz" source_b="356;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.5" name="3. Abschnitt der Reisesgeschichte / schwarz" source_b="359;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.6" name="4. Abschnitt der Reisesgeschichte / schwarz" source_b="362;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.7" name="5. Abschnitt der Reisesgeschichte / schwarz" source_b="365;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.8" name="6. Abschnitt der Reisesgeschichte / schwarz" source_b="368;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
<narration_with_stimulus id="27.3.9" name="7. Abschnitt der Reisesgeschichte / schwarz" source_b="371;2"
source_a="2" average_duration="1:00" maximum_duration="1:17" />
</subtask>
</task>
```

```
<comment id="27">
Vorgeschlagene Ablaufbeschreibungen:
1. Essen kochen (Lieblingsessen)
2. Marmelade kochen
3. Schnaps brennen
4. Auto: Reifen wechseln
5. Flug ins Ausland
6. Morgens (Kinder fertig machen)
7. Haare selbst färben
8. Hochzeit (Standesamt, Kirche, Feier)
```

```
oder EIGENES THEMA
</comment>
```

...

```
</session>
```

iLex: Handling Multi-Camera Recordings

Thomas Hanke, Jakob Storz, Sven Wagner

Institute of German Sign Language and Communication of the Deaf, University of Hamburg
 {thomas.hanke,jakob.storz,sven.wagner}@sign-lang.uni-hamburg.de

Abstract

Until recently, sign language researchers were quite happy with just one or two views for each recording session. New corpus projects, however, offer the transcriber five or more camera views. This requires much more flexibility in the transcription environment for switching between different views in order to save network bandwidth, local CPU usage, and screen real estate. Here we present a user interface study within the iLex transcription environment that allows flexible switching between video layouts whenever the transcription focus changes. Switching (including zooming) may be initiated by the user at any point of time, or can be automated to depend on tagging such as tasks or turns. The user interface is backed up by a server infrastructure providing the videos in different spatial resolutions as needed for optimal display.

1. Introduction

More than 15 years ago, we introduced the first sign language transcription environment working with digital video (syncWRITER, cf. Hanke&Prillwitz 1995). However, back then digital video in very small spatial resolution was good enough to show the video in combination with the transcript, but not really to transcribe every detail from it. Rather, one had to use VCRs – either remote-controlled by the transcription environment or directly operated by the transcriber. In the following years, technological advances finally allowed to digitize video full-size SD and then to create digital video directly with the camera and to easily transfer the material to the computer. Now, processing speed and storage capacities would also allow HD videos to be used full-size in a transcription environment. However, even on very large screens, video competes with the space needed for a useful transcription layout. This is even more true so with material that has been shot with multiple cameras. Two of our projects, Dicta-Sign and DGS Corpus, use seven cameras to record a pair of informants, too much to be displayed full-size at the same time. Sign language transcription environments such as ELAN (Crasborn&Sloetjes 2008) or iLex (Hanke&Storz 2008) have been designed at times when researchers were using digital video in the size of up to half SD (such as 320x240) and certainly need to be improved for the requirements of today's projects delivering multi-camera HD material.

ELAN allows the user to relate several media files to a transcript and to sync them. iLex just allows one single media container and relies on the container format, such as QuickTime, to group and sync several video streams into one container.

To save screen real estate, both systems allow the user to vary the display size from a fraction of the videos' spatial resolution to full size (and beyond) for all visible videos. iLex in addition allows the user to switch on or off individual tracks within the media container. This works quite fine with two or three different views grouped, but fails to provide an adequate solution when more camera views are available: A spatial layout of the tracks (defined in the container) that might be optimal when focussing on one

informant can be far from optimal in situations where both informants need to be watched in parallel.¹

In both systems, different display sizes for individual video views are not possible except by relying heavily on container formats to include one video in multiple sizes and the user switching one on and the others off as needed or to produce copies of one movie in several spatial resolutions.

Zooming onto specific parts of a video is also not possible except by providing the zoomed version as a separate movie (cf. Crasborn & Zwitserlood 2008).

Here we present a user interface study that promises to deliver the flexibility needed and at the same time to save transfer bandwidth and local processing power which even nowadays are an issue when dealing with several HD videos in parallel.

2. Screen Layouts

In our projects, transcribers have screens with native resolutions of either 1920x1200 or 2560x1440. So except for very rare cases, full HD resolution (1920x1080) is not used for transcription as the movie would occupy a good part of the screen. Depending on what they transcribe, we expect users to work more with $\frac{1}{3}$ of full HD (640x360), $\frac{1}{4}$ (480x270) or even $\frac{1}{6}$ (320x180) rather than with $\frac{1}{2}$ (960x540).²

Based on the type of discourse to be described as well as personal preferences, we expect most transcribers to work with one or two movies at a time, optionally with thumbnail-size view (160x90) for the other cameras.

2.1 Focus on one movie at a time

In this layout, clicking on any (movie or still³) thumbnail zooms the video shown so far out into a thumbnail and the thumbnail video in to the current large size. When needed, a context menu allows to switch to a two-large-movies layout.

¹ In ELAN, switching a video on or off could be easily realised from the transcript if it is the only video in its container. The layout of the videos, however, can only be influenced with respect to a left-to-right order.

² Users can still resize to any in-between value they prefer. iLex uses the next higher available resolution and scales that down.

³ Stills are preferred by some users of moving images in order to reduce visual noise.

2.2 Focus on primary views for both/all informants

With two or more large-size videos shown, thumbnails are bundled to one of the large videos. A click on a thumbnail then exchanges its movie with the bundled one.

2.3 Automatic switching based on tagging

Whenever tagging is available that is a good estimator for what the transcriber will need to focus on, this tagging can be used to switch automatically between different layouts. If for example turns have already been tagged, it makes sense to have the signer in a large view and the addressee in a small view. Good approximations to manual turn tagging can hopefully be in the near future achieved automatically through image processing (cf. e.g. Efthimiou et al. on Dicta-Sign, this volume). Another source of information is knowledge about the tasks informants are currently working on, as logged by Session Director (cf. Hanke et al.: DGS Corpus and Dicta-Sign: The Hamburg Studio Setup, this volume).

Of course, thumbnail buttons remain available to either switch to secondary views (such as birds-eye views on a single informant) or to the other informant when needed.

3. Derived Views

In addition to the views available through the films actually shot during the data collection, some derived formats are useful for the transcriber. Top of the list with HD sources certainly is zooming onto particular parts of the video, such as the signer's face. In the beginning, we ask the user to draw a frame around the signer's face. This may have to be repeated for several points in time in the video, whenever the signer moves significantly. In the future, we hope to automate this windowing through image processing (cf. Collet et al., this volume, on interfaces between transcription environments and image processing). Other examples for derived views include results of image processing such as stereo pictures.

Changes in spatial or temporal resolution alone are not considered derived views. We try to give the users the impression that any view can be scaled continuously; therefore resolution pyramids are not immediately visible to the user. As we do not see any need at this point of time to work with reductions in temporal resolution (in fact we would like to have higher resolutions available), such reductions are simply not offered as options.

We are still experimenting how to handle cropping (cutting away border stripes of the image). The idea with cropping is that anything lying outside the marked area is of no interest for transcription, and therefore the cropped movie could replace the original for all further processing. One of the problems is who might be authorised to apply cropping, as all information outside the cropped area would no longer be visible to any transcriber so errors in cropping might pass undetected.

While results of image processing might not immediately become available to the transcriber, zooms are available to the user at the click of a button: iLex just loads a higher-resolution version of the movie and then lets QuickTime crop the image in memory to the part the user is interested in. If such a derived view is used over a longer period of time, iLex marks this view to be produced as a stand-alone movie to save bandwidth and computing power on the client's side.

4. Video Server Infrastructure

Our video server currently consists of three machines with 16 processors each, attached to a SAN with a storage capacity of 100 TBytes. Two thirds of the capacity is reserved for the original footage, one third is available for caching resolution pyramids and other derived video. However, no real caching strategy is in place at this point in time. Instead, cache movies are produced as processing capacity allows. iLex then keeps track of their usage, but purging is currently left to the administrators. Our idea is to observe the system for some time before implementing strategies how to manage cache size. In the current iLex structure which allows the user to copy movies onto the local harddisk in order to work at locations where bandwidth does not allow video server access purging might render local copies useless as iLex would no longer look for them once the database entries are deleted.

Another option for the future is to provide zooming on the server side in real-time. As we currently do this on the client side, we know it can be done in real-time. Implementation on the server side, however, requires much more work, so we will first observe how much this feature will actually be used.

5. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135 and from the German Academies of Science Programme.

6. References

- Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris*, pp. 39-43.
- Crasborn, O. & Zwitterlood, I. (2008). The Corpus NGT: an online corpus for professionals and laymen. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris*, pp. 44-49.
- Hanke, T. & Prillwitz, S. (1995). SyncWriter. Integrating Video into the Transcription and Analysis of Sign Language. In T. Schermer & H. Bos (Eds.), *Proceedings of the Fourth European Congress on Sign Language Research, Munich, September, 1994*. Signum: Hamburg, pp. 303-312.
- Hanke, T. & Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages. ELRA, Paris*, pp. 64-67.

Building a Database while Considering Research Ethics in Sign Language Communities

Julie A. Hochgesang¹, Pedro Pascual Villanueva², Gaurav Mathur¹, Diane Lillo-Martin²

¹Gallaudet University, 800 Florida Avenue NE, Washington, DC 20002

²University of Connecticut, Storrs, CT 06269

E-mail: julie.hochgesang@gallaudet.edu, pedrodpedro@gmail.com,
gaurav.mathur@gallaudet.edu, lillo.martin@uconn.edu

Abstract

We are constructing an American Sign Language ID-gloss Database, which will enable sign language researchers and Deaf community members to access standard glosses for common signs. Since we are working with a language used by a community that has historically been marginalized during the research process, we feel the need to include an ethical framework for working with the Sign Language community as we consider best practices for developing sign language corpora. We will refer to the guidelines, Sign Language Communities' Terms of Reference (SLCTR), outlined in Harris, Holmes & Mertens (2009). Before making the database available to the ASL community, we plan to evaluate how members will use it and what they need from the research team to facilitate such use. This evaluation will go a long way towards ensuring that ownership of the research data lies with the ASL community. Such a reflexive evaluation of ethical practices is crucial from the beginning stages and throughout the research process. This means the ASL community is directly involved in the research process, is able to access aspects of the entire process, and can have a hand in the construction of knowledge about their own language, community and culture.

1. Introduction

We are constructing an American Sign Language ID-gloss Database, which will enable sign language researchers and Deaf community members to access standard glosses for common signs, as found in corpora such as those we are currently building. Our aim is to create a database which is flexible and powerful enough to be used by people in varying fields (e.g., linguistics, language teaching, interpreter training, preservation of Deaf heritage, etc.). As we start our work, we wish to consider not only the technical aspects of the endeavor (e.g., database design, transcription decisions, representative issues) but the ethical ones as well. We are working with a language that is used by a community that has historically been marginalized during the research process (Harris, Holmes and Mertens, 2009). It is established in spoken language corpora work that researchers need to be reflexive of ethical issues from the planning stage to publication and to be explicit about this process (Dwyer, 2006). As we consider best practices for developing sign language corpora, we feel it is necessary to also consider ethical frameworks for working with the Sign Language community. With this in mind, we are using the guidelines, Sign Language Communities Terms' of Reference (SLCTR), outlined in Harris, Holmes & Mertens (2009). This framework emphasizes "the need for the researchers to establish trust with the participants in the community and to ensure that the participants view the research as collaborative and culturally valued" (pp. 107).

2. Background – ID-gloss Database

For optimal usability, the corpora of sign languages should make data more accessible and useful; provide comprehensive and robust features for querying data;

and be in a format that is automatically searchable and retrievable. Different uses require different levels of detail in transcription, but all require consistency in notation. For this reason, we have chosen to represent signs in our corpora using ID glosses, written English words which stand for sign lemmata (Johnston, 2008; see also section 4.1 below). In order to achieve the goal of transcription using consistent ID glosses, we need a common set of sign-gloss correspondences, easily searchable, accessible, and understandable. For this reason, we are constructing an ID-gloss Database (Alkoby et al. to appear).

The ASL ID-gloss Database will consist of two main components. The first is the 'global site', which contains a pool of video files and database field templates (such as those used to describe the sign's gloss, alternative uses, morpho-syntactic category, phonological descriptions, etc.). The second component consists of multiple 'local sites', in which user groups store their own group's information about each video file, organized according to the templates chosen by that group. Due to the structure of the database, each user group has the independent ability to determine how best to structure the glosses used by that group, and which information to include in addition to the gloss itself. Furthermore, the program will allow users to see (but not modify) the glosses used by other user groups. In this way, users may choose to adopt conventions followed by other groups, possibly leading eventually to a greater degree of consistency across research groups within the United States.

The first local site will contain the glosses and additional information used by the group of Deaf and signing hearing researchers developing this project, including (in alphabetical order) Karen Alkoby, Jeffrey Bernath, Paul Dudis, Julie Hochgesang, Diane Lillo-Martin, Gaurav Mathur, Gene Mirus, and Pedro Pascual Villanueva.

3. Sign Language Communities' Terms of Reference (SLCTR)

The set of SLCTR principles is unique in that it is among the first attempts to formally draft principles towards ethical conduct for research regarding the Deaf community. While most researchers working with the Sign Language community in the past may have been mindful of how they worked with the research subjects, there has been no consistent set of principles specific to the Deaf Community that could be used by the researchers. In other words, general research ethics tend not to take into consideration specific research ethics for certain communities, including the Sign Language community. Such a lack, Harris, Holmes & Mertens (2009) claim, has led to a lack of awareness of the particular cultural issues of the Sign Language community which sometimes subsequently results in harm to the Deaf community and therefore a reluctance in the Deaf community to further collaborate with researchers. In response to this, Harris, Holmes & Mertens drafted guidelines, adapted from the Indigenous Terms of Reference (Osborne and McPhee, 2000), in order to indicate respect for, show sensitivity to, address the importance of culturally appropriate research guidelines for, and acknowledge the cultural complexity of the Sign Language community. The guidelines are reproduced in Table 1 below.

-
1. The authority for the construction of meanings and knowledge within the Sign Language community rests with the community's members.
 2. Investigators should acknowledge that Sign Language community members have the right to have those things that they value to be fully considered in all interactions.
 3. Investigators should take into account the worldviews of the Sign Language community in all negotiations or dealings that impact on the community's members.
 4. In the application of Sign Language communities' terms of reference, investigators should recognize the diverse experiences, understandings, and way of life (in sign language societies) that reflect their contemporary cultures.
 5. Investigators should ensure that the views and perceptions of the critical reference group (the sign language group) is reflected in any process of validating and evaluating the extent to which Sign Language communities' terms of reference have been taken into account.
 6. Investigators should negotiate within and among sign language groups to establish appropriate processes to consider and determine the criteria for deciding how to meet cultural imperatives, social needs, and priorities.
-

Table 1: Sign Language Communities Terms of Reference Principles (Harris, Holmes, & Mertens 2009)

4. Issues Related to our Project

As we begin work on the ASL ID-gloss Database Project, we have started to consider the project-specific issues that may arise throughout the course of our work. The three that we identify in this short paper are decisions related to gloss standardization, uses of the ASL ID-gloss Database, and transparency. We discuss each in turn in the following subsections. In general, we share the opinion that ... "the formation of partnerships with researchers and the Sign Language communities is an important step in addressing methodological questions in research" (Harris, Holmes & Mertens, 2009, pp. 111). This guides our proposed solutions, aided by the SLCTR principles, to the issues discussed in the following subsections.

4.1 Glosses – Who Decides?

Glosses are the written representations of signs using the dominant spoken language of the Sign Language community. For example, in the United States, English is used in glossing ASL. There are problems related to glossing of Sign Language data, including inconsistency and incompleteness of representations (e.g., Johnston, 2008; 1991; Slobin, 2008; Mulrooney, 2006; Pizzuto and Pietandrea, 2001), yet the practice persists. Some linguists (e.g., Johnston, 2008; 2001; 1991) propose the use of ID glosses, consistent and unique labels for signs, to take some steps toward alleviating the well-documented problems associated with traditional glosses. We agree with this proposal and have begun to establish a database in which we will maintain a catalog of ASL glosses for the research community. As we undertake this project, we are fully aware that the data we work with comes from the ASL community. We feel we have a responsibility to consult the community while constructing written representations for signs from their own language.

Principle one of the SLCTR holds that "the authority for the construction of meanings and knowledge" rests with the Sign Language community. In that vein, we plan to survey community members in determining the ID glosses included in the database. Input from the community members will help to establish the optimal gloss we will use for each sign. We will target members of different sub-communities, including those with different backgrounds and those with different possible uses of the database (cf. section 4.2) in order to get a representative response.

Principle 5, in which the complexity of the cultural make-up of the Sign Language community is considered during the research process, is inherent in our treatment of the glosses as equal representations of as many ASL signs as we can feasibly include. Variation based on region, age, gender, education and other social factors will not be used to include or exclude any certain ASL sign. If the signs are linguistically different (based on our ultimate set of criteria), they will receive different

ID glosses. We will not intentionally exclude signs that may be considered by some to be used by a minority of the Sign Language community. In this treatment of the data, we avoid highlighting certain ASL signs as representative of the entire Sign Language community. We will stress in the literature regarding our database that any unintentional exclusion is due to our being unaware of such signs, as well as our limitations by time and funding to including only a subset of all signs.

We are also mindful of the fact that glosses are not cultural artifacts (as pointed out by our collaborator Paul Dudis) but tools of the scientific realm. This means that ultimately factors including the goals of the research project, the issues well discussed in the field regarding glosses and representation of data, and the input from the Sign Language community will all be considered as we make our final decisions in selecting the ID glosses to be used in our component of the database. All of the factors discussed here have also entered into our decisions regarding the design of the database, and in particular our implementation of a system which will allow different user groups to construct their own catalog of ID glosses which are best suited for their own purposes.

4.2 How the Database Will Be Used

The Amsterdam Manifesto, prepared by a group of sign linguists following the meeting of the conference on Theoretical Issues in Sign Language Research in Amsterdam in 2000, raises the point that much of sign language research is dependent on Deaf research assistants as well as data from Deaf native signers. The manifesto suggests that one way to acknowledge the contributions from these sign language communities is to give something back to them.

The ID-gloss database as described above clearly draws on and describes data from Sign Language community members. The question raised by the Amsterdam Manifesto and SLCTR regarding the database is, then, what can the investigators give back to the Sign Language communities in exchange for establishing this database? Is it sufficient to allow access to the database by the Sign Language community members? These questions ultimately depend on the issue of how the database is to be used.

The second and sixth principles of the SLCTR provide guidance in addressing these concerns. In their discussion of the second principle, Harris, Holmes & Mertens (2009) talk about how important it is to publish some of the research in sign language, rather than publishing in written language all the time. The underlying premise of this principle is that Sign Language community members should have access to the research, and publishing some of the work in sign language is one way to provide that access. The sixth principle says, in essence, that investigators should work

with Sign Language groups to establish processes so that the research would meet the Sign Language communities' priorities. These principles can be applied in the context of the ID-gloss database. Here, we outline two ways that we do this.

First, we make the database as accessible as possible to the Sign Language community members. It is important to bear in mind that the database is intended to be a research tool that enables easier and more consistent transcription. It is not intended to be a dictionary, even though it shares some elements in common with one (e.g., an entry will include an image of the sign, a corresponding gloss, its meaning and its phonological description, among others). However, this intended use does not mean that we cannot share the database with Sign Language community members, and that they would not find appropriate uses for it. We could, for example, design a user interface specifically for Sign Language community members that would permit them to understand clearly the purpose of the ID-gloss database. This would address the second principle, in which we acknowledge their right to ensure that what they culturally value as a Sign Language community is included.

Another way to address the sixth principle is to set up guiding principles, in close consultation with Sign Language community members, on how to use the database. The guiding principles should clarify, for example, whether users are allowed to download and/or disseminate the information from the database. The guidelines should also specify who can add and modify entries in the database, and for what purposes the database can be used, e.g., for a conference presentation, for classroom instruction, and/or for purely research purposes.

By opening up the ID-gloss database to Sign Language community members, issues of ownership and researchers giving something back to the community are at least partially addressed.

4.3 Transparency

Transparency requires that researchers are open and reflexive about their information regarding the community being studied. In terms of the Sign Language community, researchers must adhere to transparency in a way that is accessible, i.e., in the community members' own sign language. Being transparent is a factor in meeting most of the SLCTR principles.

On the website where our ASL ID-gloss Database is hosted, we will provide signed ASL text wherever there is written English text. This practice of providing Sign Language text has been established by some other signed language corpora (e.g., the BSL corpus which can be found at: www.bslcorpusproject.org, last accessed March 20, 2010). We intend to adopt this

practice. In addition, the specialized user interface, as introduced in section 4.2, should allow Sign Language community members opportunities to provide feedback on aspects of the database, e.g., through comment boxes that accept video media (therefore signed input) and through polling. This is directly concerned with the sixth principle of the SLCTR, in which the Sign Language community helps establish research procedures. We will provide community members with the accessible opportunity to give input on the design and content of the database in a way that reflects their priorities.

By being transparent, we indicate our respect and understanding for practices culturally appropriate to the ASL community.

5. Discussion

We would like to emphasize that while we deem it extremely important that the Sign Language community be involved in the research process, we are aware that they do not possess the same scientific training or knowledge as sign language linguists do. We plan to honor the SLCTR, Amsterdam Manifesto, and the Sign Language community by being reflexive of and transparent about our practices and collaborating with the Sign Language Community, while simultaneously meeting the requirements of the research community. In fact, the membership of the Sign Language community and the research community overlaps, as there are some sign language linguists who are Deaf or otherwise members of the Sign Language community; there are of course also some sign language linguists who are not members of the Sign Language community. The SLCTR principles apply equally to all sign language researchers.

In this paper, we have discussed a few particular strategies regarding how we are implementing the SLCTR principles, including our actions and which SLCTR principles they reflected. We plan to continue consulting the SLCTR, including the principles we did not address in this paper, throughout the process of our research project.

6. Conclusion

As researchers, our focus is usually on theoretical, experimental, and/or technical aspects of our projects. However, it is important for us to bear in mind that the language we are so deeply involved in studying has a rich and important cultural value to the members of the Sign Language community. To appropriately follow relevant ethical considerations as we conduct our research, we must consciously consider and implement principles which have been determined to be suitable and applicable for studies in this area. Such a reflexive evaluation of ethical practices is crucial from the beginning stages and throughout the research process. This means the Sign Language community, in our case the ASL community, is directly involved in the research process, is able to access aspects of the entire process,

and can have a hand in the construction of knowledge about their own language, community and culture.

7. Acknowledgements

The project described in this article was supported by Award Number R01DC009263 from the National Institute on Deafness and Other Communication Disorders. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Deafness and Other Communication Disorders or the National Institutes of Health.

Thanks to Paul Dudis and Raychelle Harris for input on this article.

8. References

- Alkoby, K., Bernath, J., Hochgesang, J.A., Mirus, G., & Pascual Villaneuva, P. (to appear). Construction of an ID Gloss Database for American Sign Language. To be presented at the conference on *Theoretical Issues in Sign Language Research*, Purdue University, October 2010.
- Dwyer, A.M. (2006). Ethics and Practicalities of Cooperative Fieldwork and Analysis. In J. Gippert, N.P. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation*. Berlin: Mouton de Gruyter, pp. 31--66.
- Harris, R., Holmes, H.M., & Mertens, D.M. (2009). Research Ethics in Sign Language Communities. *Sign Language Studies*, 9(2), pp. 104--131.
- Johnston, T. (2008). Corpus Linguistics and Signed Languages: No Lemmata, No Corpus. In O. Crasborn, E. Efthimiou, T. Hanke, E.D. Thoutenhoofd, I. Zwitserlood (Eds.), *Proceedings of the Third Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, pp. 82--87.
- Johnston, T. (2001). The Lexical Database of Auslan (Australian Sign Language). *Sign Language and Linguistics*, 4(1/2), pp. 145--169.
- Johnston, T. (1991). Transcription and Glossing of Sign Language Text: Examples from Auslan (Australian Sign Language). *International Journal of Sign Linguistics*, 2(1), pp. 3--28.
- Mulrooney, K. (2006). *The Structure of Personal Narratives in American Sign Language*. Doctoral dissertation. Gallaudet University: Washington, DC.
- Osborn, R. & McPhee, R. (2000). Indigenous Terms of Reference (ITR). Presented at the 6th UNESCO-ACEID International Conference on Education, Bangkok, December 12-15.
- Pizzuto, E & Pietandrea, P. (2001). The Notation of Signed Texts: Open Questions and Indications for Further Research. *Sign Language and Linguistics*, 4(1/2), pp. 29--45.
- Rathmann, C., Mathur, G. & Boudreault, P. (2000). Amsterdam Manifesto. *Das Zeichen* 14, pp. 654--655.
- Slobin, D. (2008). Breaking the Molds: Signed Languages and the Nature of Human Language. *Sign Language Studies*, 8(2), pp. 114--130.

Development of a Moodle VLE Plug-in to Support Simultaneous Visualisation of a Collection of Multi-Media Sign Language Objects

Markus Hofmann¹, Kyle Goslin², Brian Nolan³, Lorraine Leeson⁴, Haaris Sheikh⁵

^{1,2,3} Department of Informatics
Institute of Technology Blanchardstown
Blanchardstown Road North
Dublin 15
Ireland

⁴ Centre for Deaf Studies
40 Lower Drumcondra Rd.
Drumcondra
Dublin 9
Ireland

⁵ Interesource Group (Ireland) Limited
48 South William Street
Dublin 2
Ireland

E-mail: markus.hofmann@itb.ie, kylegoslin@gmail.com, brian.nolan@itb.ie, leesonl@tcd.ie, haaris@interesourcegroup.com

Abstract

Using Virtual Learning Environments (VLE) to support blended learning is very common in educational institutes. Delivering learning material in a flexible and semi-structured manner to the learner transforms such systems into powerful eLearning tools. However, the presentation and visualisation of individual or multiple learning objects is mostly dictated by the system and cannot be altered easily.

This paper reports on a project between Trinity College Dublin (TCD) and the Institute of Technology Blanchardstown (ITB) that aims to improve the simultaneous visualisation of multiple multimedia objects for deaf learners of ISL. The project was implemented using the Open Source VLE Moodle. Moodle's nature of being Open Source and having the ability to code plug-ins qualified it to be the most suited vehicle to address the visualisation problem. Traditionally VLEs allow the viewing of one learning object at a time, which meant that deaf learners could either view a pre-recorded, signed in ISL, video lecture or concentrate on textual accompanying content but not both. The developed Moodle plug-in allows academics to group multiple videos into a 'lecture'. It further facilitates the addition of rich text content to each video. The learner can select and view one video from a possible sequence of many as well as view the text that belongs to the video. The paper further outlines detailed implementation and techniques applied.

1. Introduction

Using Virtual Learning Environments (VLE) to support blended learning is very common in educational institutes. Delivering learning material in a flexible and semi-structured manner to the learner transforms such systems into powerful eLearning tools. However, the presentation and visualisation of individual or multiple learning objects is mostly dictated by the system and cannot be altered easily.

Irish Sign Language (ISL) is an indigenous language of Ireland and is recognized by the EU as a natural language. It is a language separate from the other languages used in Ireland, including Irish, English and,

in Northern Ireland, British Sign Language. Some 6,500 Deaf people use ISL on the island of Ireland. Our goal is to deliver third level programmes to students online to resolve problems of time, geography and access, maximizing multi-functional uses of digital assets across our programmes to maximize the "Deaf-friendliness" of blended learning delivery for Deaf and hard of hearing students.

This paper reports on a project between Trinity College Dublin (TCD) and the Institute of Technology Blanchardstown (ITB) that aims to improve the simultaneous visualisation of multiple multimedia objects for deaf learners of ISL. The project was

implemented using the Open Source VLE Moodle. Moodle's nature of being Open Source and having the ability to code plug-ins qualified it to be the most suited vehicle to address the visualisation problem. Traditionally VLEs allow the viewing of one learning object at a time, which meant that deaf learners could either view a pre-recorded, signed ISL, video lecture or concentrate on textual content but not both. The developed Moodle plug-in allows academics to group multiple videos into a 'lesson'. It further facilitates the addition of rich text content to each video. The learner can select and view one video from a possible sequence of many as well as view the text that belongs to the video. The paper further outlines detailed implementation and techniques applied.

2. Background

2.1 Deaf Studies in Ireland and Europe

Approximately 1 person in a 1000 is a signed language user (Johnston 2004, Conama 2008), which suggests that there are some 490,426 Deaf signed language users in the EU. In Ireland, there are approximately 5,000 Irish Sign Language users in the Republic (Matthews 1996) and an approximate 1,500 ISL users in Northern Ireland. Irish Sign Language (ISL), an indigenous language of Ireland, is recognized by the European Union as a natural language. It is a language separate from the other languages used in Ireland, including English, Irish, and, in Northern Ireland, British Sign Language. Some 6,500 Deaf people use ISL across the island of Ireland. In great part, because of the history of suppression of signed languages across the EU, the average Deaf person leaves school with a reading age of 8.5 to 9 years. Given this, it is no surprise that Deaf people are the most under-represented of all disadvantaged groups at third level. This poses two initial challenges: (1) getting Deaf people into third level and (2) presenting education in an accessible form (Nolan and Leeson, 2009).

In tackling these challenges, Trinity College Dublin and the Institute for Technology, Blanchardstown, Dublin (ITB) have partnered to create a unique eLearning environment based on Moodle as the learning management system, in the delivery of Deaf Studies programmes at TCD. This partnership delivers third level programmes to students in a way that resolves problems of time, geography and access, maximizing multi-functional uses of digital assets across our programmes. Our digital assets include a corpus of ISL, the 'Signs of Ireland Corpus' which is one of the largest, most richly annotated in the world. We have operated with some online delivery since 2005, hosted by ITB, and in early 2008 were successful in attracting significant Irish government funding to expand delivery of a series of undergraduate diplomas to degree level nationwide under the Strategic Innovation Fund, Cycle II and SIGNALL II.

2.2 Moodle VLE

Moodle is a popular open-source course management system that can be scaled from several users and courses to several hundred thousand users with thousands of course modules. The VLE is used around the world and is available in approximately 100 different languages. Moodle has almost 50,000 validated installations in over 200 countries with a total of over 30 million users (Moodle, 2010).

One of the strength of this VLE is that, firstly, it is open-source which makes it possible to access and change the code but also, and more importantly, that the framework Moodle is based on allows the development of plug-ins that easily slot into the existing structure of the application. This project takes advantage of this framework and developed the plug-in that lets users view video and text side by side to support the learning requirement of deaf learners.

2.3 Project Rationale

Signed languages, by their nature, are visual-gestural languages, which (unlike spoken languages) do not have a written form. Given this, the online content is required to be multi-modal in nature and we utilize rich-media learning objects in our delivery. This presents a number of serious and important challenges which include:

- Universal design in an online curriculum for Deaf and hearing students;
- Identifying what aspects of ISL learning can best be supported & assessed online;
- Assessing signed language interpreting skill in an online context;
- Decisions regarding ISL annotation & mark-up standards;
- Using the Signs of Ireland corpus in blended learning contexts;
- Leveraging a corpus within digital learning objects in a Moodle environment;
- Architecture of a digital learning environment to support ISL learning;
- Issues of assessment in an eLearning context;
- Creating a plug-in for Moodle to facilitate delivery of large multimedia files online rather than text-only data.

2.4 Learning Objects

The learning objects that are of significance to the deaf learner are signed video recordings that are accompanied by text. It was also of importance that the structure of several videos and their textual content can be incorporated into the structure of how Moodle presents these learning objects.

3. Moodle Plug-in Development

Moodle offers a complete plug-in framework that allows developers to create custom learning resource containers that are fully integrated into the VLE application. In particular the complete integration into Moodle is of crucial importance so that existing Moodle environments and functionality can be fully utilised without having to

separately re-develop them. This decreases project development time as well as cost of the overall project considerably. Once developed the resource container becomes an option when adding a new learning object to a course page. When a custom resource container is created by a developer, the editing and viewing of this container's styles and layouts can be completely decided by the developer. This openness and flexibility of the resource container implementation was suited for the development of this project. A resource plug-in was then developed for the Moodle platform as a way to deliver the content of the Deaf Studies Lesson.

The plug-in developed as part of this project is an Activity Module that allows lecturers to create a learning object called 'Lesson'. Once created, each lesson can be populated with one or more videos and each video can be associated with format rich textual information including embedded images and links. Several aspects needed to be considered:

- The plug-in needed to work independently of Moodle settings and should not depend on any other optional modules or plug-ins. This was important so that it works for all Moodle installations without setting any system requirements other than the version number (version 1.9).
- The plug-in takes advantage of the xmlDB framework which facilitates the creation and manipulation of new data tables that are part of the main Moodle database. Each video therefore creates a new record in a table that needs to be included in the backup functions that Moodle has as core functionality. This was solved using specific backup functions that included the data created by the plug-in to be incorporated in the overall (and pre-existing) backup procedure provided by Moodle.
- This tool will also be used as part of a European project which meant that localisation was required resulting in the requirement of the integration of different languages is a possibility. Certain domain

specific terminology is not part of Moodle's language packs and it was therefore necessary to include an environment that facilitated multi-language support. The selected default language of the Moodle installation is also the default language of the developed plug-in. Should the plug-in language pack for the selected default language be unavailable, it reverts back to English as second default language. Additional languages can be added easily by modifying one single file.

- Moodle constantly stores users' activities recording IP address, date, time, and viewed learning object. Traditionally most learning objects uploaded and displayed by Moodle are single files. Our tool however produces a learning object called 'Lesson' that consists of several files (one for each video and text). This meant that data collection needed to be integrated in such a way that the Moodle log files also include the users' behaviour within the 'Lesson' object. In brief, we wanted detailed user data in relation to the selection of videos. We believe that this data will provide us with information that could be used for subsequent analyses.
- The last consideration was the availability of online, on the spot help functionality. Ambiguous plug-in sections as well as text boxes now have a help button that describes the rationale behind the object and what values it expects.

4. Features of the Visualisation Tool

The tool contains a number of unique features as outlined in the following subsections.

4.1 Simultaneous Visualisation

The key feature of the tool is the facility to display formatted information (including rich text format, links and images) and video side by side to improve the content provision for deaf students. This is an important aspect considering the different learning characteristics

The screenshot shows a Moodle course page titled "Introduction Video 4". The page layout includes a Moodle logo at the top left, a navigation breadcrumb "learning space > Sign Ling > Lecture 1 - Introduction", and a "Jump to..." search box at the top right. Below the breadcrumb, the video title "Introduction Video 4" is displayed, along with a "Video Number" selector showing options 1, 2, 3, and 4 (with 4 selected). The main content area is split into two columns. The left column contains text about Moodle, stating it is a popular open-source course management system used in 100+ languages by 50,000+ installations in 200+ countries, with 30 million users. It also mentions the VLE's strength as open-source. The right column features a video player showing a woman signing. The video player has a "jwplayer" logo and a progress bar at the bottom. At the very bottom of the page, the number "118" is displayed on the left, and a "Previous Video" link is on the right.

of deaf learners. This is displayed in Figure 1 showing the layout of the tool. Learners can navigate either directly to the desired video by clicking on the ‘Video Number’ or use the ‘Next’ and ‘Previous’ links located underneath the video screen. The text area to the left of the video allows the lecturer to add subject related text, images and hyperlinks. In case the content exceed the space provided a scrollbar will allow the student to move up and down.

4.2 Localisation

The plug-in will be used in various different countries across Europe and therefore a great importance was to add the ability for the tool’s interface descriptions strings to be localised to the users native language. The project uses the plug-in framework provided by Moodle which makes it possible to take advantage of the native localisation features that are available to all the components of the VLE. Localised terms are added to the language files inside the plug-in and whatever language is selected by the administrator for the installation of Moodle, will reflect which one of the languages the plug-in interface description strings will be translated to. If the terms are not available in the respective language files the default English values will be applied.

4.2 Data Collection

When implemented on a live server, the students who interact with the tool provide a wide range of unique personal usage patterns, that provide an insight to the most used and disused sections of the tool. For this reason, data collection facilities were integrated into the tool to capture this data.

Two data collection tools were used to record this data. The first of these was Moodle’s integrated data collection tool. By default Moodle stores information about the various different pages which were accessed by students into its database. This information when filtered, gives an initial raw look at the most visited parts of the plug-in. In addition to this a data collection tool was created that records plug-in specific data such as which video has been watched and for how long. This data collection feature therefore lets the educator analyse the click stream of each individual user creating a better picture of the learners. More frequently watched video could indicate that the topic is particularly difficult to understand.

4.2 Structure and Administration

The deaf studies content visualisation tool can be managed within Moodle and it follows the same structure and layout constraints as the native Moodle learning objects. Keeping this uniformed creation and editing process was of great importance, to ease the learning curve of the tool so administrators do not have to learn a new procedure but can apply their existing Moodle knowledge to create lessons and their respective video and text content. When adding new videos to the

lesson which was created, various different content manipulation controls are available to the administrator. These controls give the administrator the ability to move the current video further up or down on the list of videos for the lesson, delete videos from the lesson and also edit an individual lesson sections with a full HTML markup editor.

5. Future Process & Development

5.1 Implementation

At present the tool has passed its piloting stage and will be rolled out in the coming weeks in countries such as Ireland, UK, Poland, Czech Republic and Finland. The feedback of this rollout will be used to improve and finalise the tool after which it will be provided as downloadable plug-in under the GNU General Public Licence.

5.2 Future Development

One of the future development aims is to add multiple video tagging so that the textual content assigned to a video can change based on the progression of the video clip. This is of particular importance when longer videos are used as learning objects. Existing work reviewing video tagging tools, and investigating automated segmentation (for example see Campos et al. 2008), suggests that it will be possible to add such features to our tool.

In addition we aim to improve the visualisation of the student’s usage pattern data to give the teacher a graphical representation of the most beneficial aspects of the course content which was provided

5.3 Data Analysis

The tool described in this paper has extensive data collection facilities as described above. These data can be used to gain previously unknown patterns and learner behaviour. For example, it will be possible to investigate whether there is a correlation between usage of the tool and academic performance. Other interesting measures could be students’ time, frequency and duration of usage.

6. Conclusion

This paper reported on the development of a Moodle VLE plug-in that offered simultaneous visualisation of a collection of multi-media sign language corpora objects. Initial feedback from students indicates the level of success of this project in terms of improved lecture content provision. In particular the option to have such a tool fully integrated in one of the most frequently used VLE adds considerable value to the plug-in. The feature list of the plug-in includes novel methods of learning object visualisation, localisation for multi-language support, data collection ability for subsequent data analyses, integrated backup solution, and an online help that is also integrated into the Moodle framework to add consistency.

7. Acknowledgements

We would like to acknowledge the funding from the Strategic Innovation Fund (Cycle 2), SIGNALL II as well as ongoing support from the Institute of Technology Blanchardstown and the Centre for Deaf Studies (TCD).

8. References

- Campos, L., Fernández-Luna, J., García, J., Gómez, F., Huete, J. and, C. Martín-Dancausa (2008). A Video Segmentation and Annotation Tool for Parliamentary Recordings and Transcriptions. IADIS International Conference Informatics. Amsterdam, Netherlands.
- Conama, J.B. (2008): Review of the Signing Information Project, Mid-West Region. Limerick: Paul Partnership.
- Johnston, T (2004) W(h)ither the deaf community? Population, genetics, and the future of Australian Sign Language, *American Annals of the Deaf* 148:5.
- Matthews, Patrick A. (1996). *The Irish Deaf Community- Volume 1*, ITE, Dublin.
- Moodle (2010). Moodle Statistics (online) <http://moodle.org/stats>. Last accessed: March 2010.
- Nolan, B. and L. Leeson (2009). Creating access to education with progression pathways via blended learning of Deaf Studies at third level in Ireland: Open innovation with digital assets. *ITB Journal*, Issue 18, pp 72-83.

Eliciting Spatial Reference for a Motion-Capture Corpus of American Sign Language Discourse

Matt Huenerfauth

The City University of New York (CUNY)
Computer Science Department, Queens College
65-30 Kissena Blvd, Flushing, NY 11367 USA
E-mail: matt@cs.qc.cuny.edu

Pengfei Lu

The City University of New York (CUNY)
Computer Science Program, Graduate Center
365 Fifth Ave, New York, NY 10016 USA
E-mail: pengfei.lu@qc.cuny.edu

Abstract

We seek computational models of the referential use of signing space and of spatially inflected verb forms for use in American Sign Language (ASL) animations for accessibility applications for deaf users. We describe our collection and annotation of an ASL motion-capture corpus to be analyzed for our research. We compare alternative prompting strategies for eliciting single-signer multi-sentential ASL discourse that maximizes the use of pronominal spatial reference yet minimizes the use of classifier predicates.

1. Introduction

Significant numbers of deaf adults in the U.S. have relatively low levels of written English literacy (Traxler, 2000); many have difficulty reading English text on websites or other information sources. Animations of American Sign Language (ASL) make information and services accessible for these individuals. There are two major types of ASL animation technologies: scripting and generation/translation software. Scripting software allows a human author to specify the movements of a virtual human character by arranging signs and facial expressions on a timeline to be performed, e.g. (Vcom3D, 2010; Kennaway et al., 2007). Generation/translation software automatically synthesizes ASL sentences, given an English input sentence to be translated; Huenerfauth and Hanson (2009) describe and review such systems.

Our goal is to construct computational models of ASL that could be used to partially automate the work of human authors using scripting software or to underlie generation/translation systems. Specifically, we wish to model aspects of ASL linguistics that are not handled by modern ASL scripting or generation software. Signers associate entities under discussion with 3D signing space locations, and signs whose paths or orientations depend on these locations pose a special challenge: They are time-consuming for users of scripting software to produce, and they are not included in the repertoire of most modern ASL generation/translation software.

Huenerfauth (2009) found that native signers' comprehension of ASL animations improved when the animations included: (1) association of entities with locations in the signing space and (2) the use of verbs whose motion paths were modified based on these locations. Thus, users of ASL animation software would benefit from better handling of these two phenomena.

Section 2 describes how these spatial reference phenomena are frequent in ASL signing and important to the meaning of ASL sentences. Section 3 describes our overall research goals of: (1) collecting an ASL corpus using motion-capture equipment and video, (2) annotating the use of spatial reference phenomena and other linguistic features in this corpus, and (3) analyzing the human movement data in this corpus (and its

relationship to the linguistic structure) to build computational models of how ASL signers associate entities under discussion with 3D signing space locations. These computational models will be incorporated into ASL animation technologies we are developing to make the resulting animations more realistic, understandable, and ultimately more useful for deaf users in accessibility applications. Section 4 discusses our corpora collection and annotation procedure; section 5 compares alternative prompting strategies we have used during year 1 of the project to elicit signing performances of the desired form. Section 6 contains conclusions and future research plans.

2. Spatial Reference Points in ASL

As in other sign languages, users of ASL frequently associate entities under discussion with locations in the signing space involved in later pronominal reference and other purposes (Liddell, 2003; Meier, 1990; Neidle et al., 2000). Various ASL constructions can be used to establish a *spatial reference point (SRP)* for some entity:

- Pre-nominal determiners and some post-noun-phrase adverbs consist of a pointing sign in which the entity in that noun phrase is assigned a 3D SRP location.
- Fingerspelling or some nouns may also be signed outside their standard location to establish an SRP.

The movements of other ASL signs are parameterized on the 3D locations of previously established SRPs:

- Personal, possessive, and reflexive pronouns consist of pointing movements to SRPs' 3D locations.
- Some verbs change their motion path or orientation to indicate the 3D location of their subject, object, or both. What features are modified and whether this is optional depends on the verb. These *inflecting verbs* (Padden, 1988) are sometimes referred to as agreeing (Cormier, 2002) or indicating verbs (Liddell, 2003).
- During verb phrases or possessive phrases, the SRP of the subject/object or possessor/possessed may be indicated by head-tilt/eye-gaze (Neidle et al., 2000).

Thus, ASL animation software that does not model SRPs cannot generate: determiners, pronouns, many noun phrases, some verb phrases, spatially inflected verbs, or possessive phrases – all of which are based on SRPs.

3. Research Goals

We seek computational models of: (1) what locations in 3D space are commonly chosen for SRPs, (2) which entities are assigned SRPs, (3) how the motion-paths of inflecting verbs change based on the 3D location of their subject’s and object’s SRP. Producing the hand, eye, and head movements needed to establish and refer to SRPs is burdensome for human users of ASL scripting software – and producing accurate 3D movements of spatially inflected ASL verbs is even harder. We believe that models of these three issues above could be used to partially automate this work or used to fully automate the work of ASL-animation generation/translation software.

We will build these computational models through the collection and analysis of a motion-capture corpus of ASL multi-sentential discourse. We hypothesize that linguistic features of the discourse affect the likelihood of a signer assigning an entity an SRP (and where it will be placed); we will analyze the corpus using statistical machine learning techniques to build SRP establishment models. We also believe that mathematical functions of verbs’ motion paths (parameterized on SRP locations) can be induced from the collected 3D motion data; we will use regression/model-fitting techniques to construct an animation lexicon of ASL inflecting verbs that are spatially parameterized on the 3D location of the subject and/or object (so that inflected forms can be synthesized as needed by ASL scripting or generation software).

This corpus consists of motion-capture recordings of multi-sentential discourse with annotation of SRP establishment and reference. Prior researchers collected video-based corpora, e.g. (Neidle et al., 2000; Bungeroth et al., 2006; Efthimiou & Fontinea, 2007), or short sign language recordings via motion-capture, e.g. (Brashear et al., 2003; Cox et al., 2002). Researchers have designed schemes for annotating the referential use of signing space (Lenseigne & Dalle, 2005), but no previous motion-capture corpus includes such SRP annotation.

4. Corpora Collection Procedure

For our corpus, we record handshape; hand location; palm orientation; eye-gaze vector; and joint angles for the wrists, elbows, shoulders, clavicle, neck, and waist. Our novel configuration of commercial motion-capture equipment includes: two Immersion CyberGloves®, an Applied Science Labs H6 head-mounted eye-tracker, an Intersense IS-900 inertial/acoustic tracker (for the head), and magnetic/inertial sensors on an Animazoo IGS-190 bodysuit. Three high-definition digital video cameras record front, side, and facial close-up views of the signer (referred to as the “performer”). Another native signer (the “prompter”) sits behind the front-view camera to converse with the performer and elicit signing to record.

In our first year, we have recorded and annotated 58 ASL passages from 6 signers (~ 40 minutes of data). To collect natural use of SRPs, we elicit *unscripted* multi-sentential single-signer discourse. Table 1 lists different prompting strategies we tried and how many recordings we collected using each. The totals for each vary because the recording session was intentionally kept relaxed/conversational to promote more natural signing:

Type	N	Description of the Prompting Strategy
Personal Intro/Info	15	Introduce yourself, describe some of your background, hobbies, family, education...
Hypothetical Scenario	4	What would you do if: You were raising a deaf child? You could have dinner with any two famous or historical figures?
Compare (not people)	9	Compare two things: e.g. Mac vs. PC, Democrats vs. Republicans, high school vs. college, Gallaudet University vs. NTID, travelling by plane vs. travelling by car, etc.
Compare (people)	7	Compare two people you know: your parents, some friends, family members, etc.
Recount Movie/Book	7	Tell us about your favorite movie or your favorite book. What happens in it?
Tell a Story (3 Wishes)	2	Invent a story using this topic: “If I had a genie that could grant three wishes, I’d...”
Repeat Conversation	6	Watch 3-minute video of ASL or captioned conversation, then explain what you saw.
Children’s Book	5	Read a short children’s book, then explain the story as you remember it.
Wikipedia Article	3	Read a 300-word Wikipedia article on “The History of Racial Segregation in the United States.” Explain/recount the article.

Table 1: Types of prompts used during data collection with the number of stories of each type collected (N).

the prompter used different strategies to elicit signing from the performer. Sometimes the performer was verbose in their response to a prompt, but other times, he/she could think of little or nothing to say. Further, since performers were recorded for only 1 hour (after the motion-capture equipment was set-up and calibrated), we rarely had sufficient time to try all of the different prompt-types during each performer’s recording session.

After collecting each story, we synchronize our video and motion-capture streams, apply the data to a 3D skeleton, and produce video segments for each story. A team of native ASL signers (including students from deaf high schools in New York) annotates the data using the SignStream™ annotation tool (Neidle et al., 2000). We annotate some traditional information: sign glosses; part-of-speech; syntactic bracketing of NPs, VPs, clauses, sentences; and non-manual marking of role shift, negation, who/what/where/when/why/how questions, yes-no questions, topicalization, conditionals, and rhetorical questions. In support of our research goals, we also annotate: when SRPs are established, which discourse entity is associated with each, when referring expressions indicate each SRP, and when any verbs are spatially inflected to indicate each SRP. Each SRP is assigned an index number, and each pronominal or verb reference to an SRP is marked with this index. These SRP establishments and references are recorded on parallel timeline tracks to the glosses and other linguistic annotations. We also mark any *classifier predicates (CPs)* performed; CPs are special signs in which the signer synthesizes a movement for the hands (or sometimes the

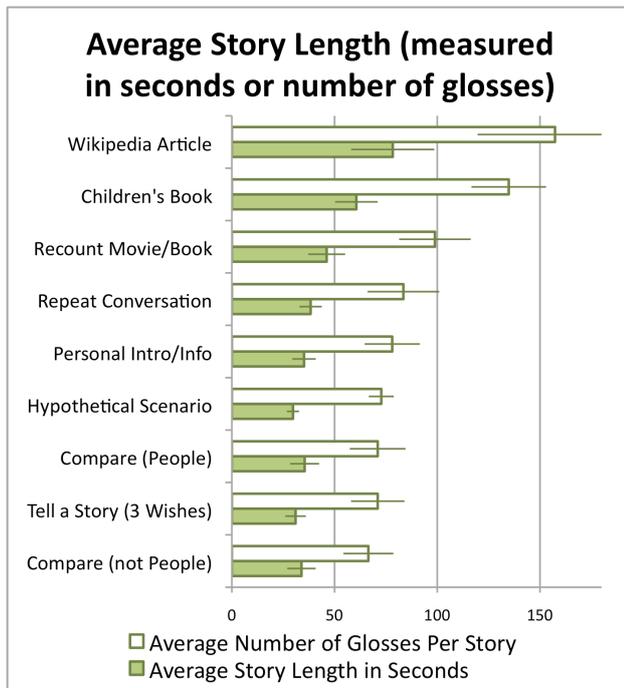


Figure 1: Length of the ASL stories collected.

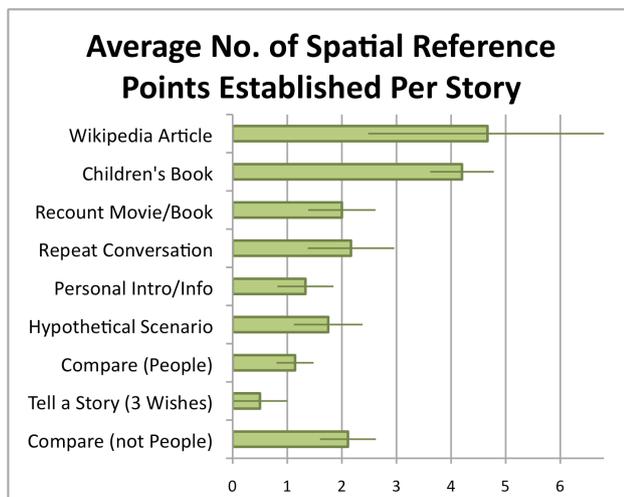


Figure 2: Spatial reference points established.

body) to indicate the spatial arrangement, size, shape, or movement of people/objects in a 3D scene being described. We count CPs in order to measure the effectiveness of our prompting strategies (see section 5).

5. Comparison of Prompting Strategies

After collecting/annotating the first 58 stories, we can determine which prompting strategies were effective at collecting the desired type of ASL signing. An ideal ASL story to be collected for this corpus would:

- Be long enough to allow for establishment of SRPs.
- Sometimes contain multiple SRPs (perhaps 3+) to enable the study of diverse types of spatial use.
- Contain as many pointing signs (determiners, pronouns, etc.) or inflected verbs that refer to SRPs as possible. With many examples of these *spatial references (SRs)*, we will be able to study diverse forms of spatial use and reference in ASL signing.

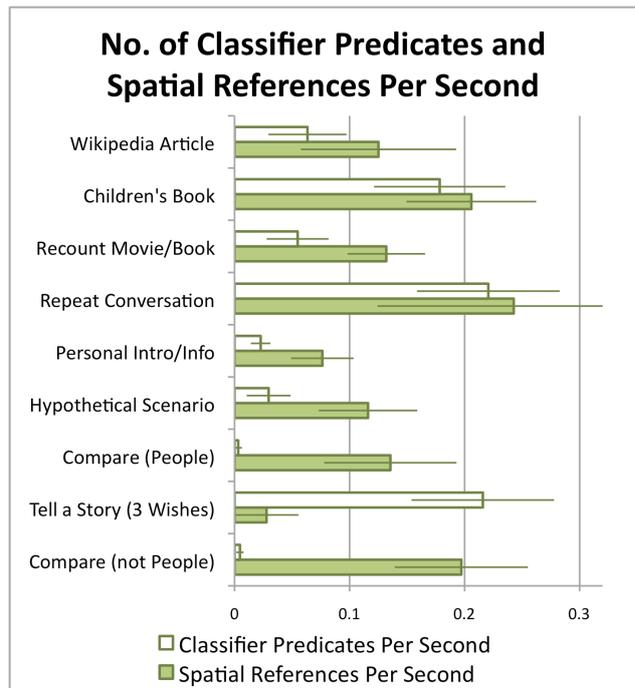


Figure 3: Number of classifier predicates and spatial references per second in each type of ASL story.

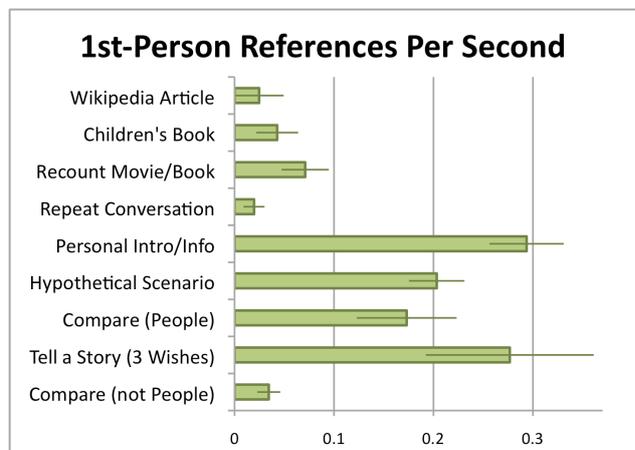


Figure 4: First-person references in the ASL stories.

D. Contain as few CPs as possible. CPs complicate how signing space is used; the interaction between CPs and SRPs is beyond the scope of our current work.

Figure 1 displays the average length of the stories collected using each prompting strategy – as measured in seconds of time or in the total number of manual signs (count of annotated glosses). Prompt types are listed in descending order based on their number of glosses; they are listed in this same order in Figures 2, 3, and 4. Error bars indicate the standard error of the mean for each value. The longest stories arose from prompts in which the performer recounted an article, book, movie, or conversation they saw recently or had seen in the past.

As listed in criterion ‘B,’ we’d like to collect some stories in which signers establish larger numbers of SRPs. Figure 2 displays the number of SRPs established in each story (entities assigned 3D locations for pronominal use). The longer stories generally contained more SRPs. (N.B.

If the performer referred to the prompter during the story, then the count of SRPs for that story was increased by 1. In such cases, the addressee was used as a 2nd-person referent, and thus, we counted the addressee as an SRP.)

Criteria ‘C’ and ‘D’ explain how we want to maximize the number of SRs in each story and minimize the number of CPs. Figure 3 displays the average frequency of SRs and CPs (as measured per second) in stories of each prompt-type; the values are displayed on the same graph to enable comparison of their ratio. The SRs in Figure 3 include 3rd-person and 2nd-person references, but not 1st-person (e.g. signs like “me”/“my” or inflecting verbs in which the subject/object is the signer) because these do not involve pointing to a location in the surrounding signing space. While we are not particularly interested in maximizing or minimizing the frequency of 1st-person references, we present their frequency in Figure 4 – for the sake of completeness. Unsurprisingly, the “personal intro/info,” “tell a story,” and “hypothetical scenario” prompts led to many 1st-person references. In some of the “compare (people)” stories, signers compared *themselves* to someone else.

6. Conclusions and Future Work

Our analysis of the different prompting strategies will guide our future data collection. Based on their high CP/SR ratio, we will no longer use the “tell a story,” “children’s book,” and “repeat conversation” prompts. The long story lengths, high number of SRPs established, and modest CP/SR ratio of the “Wikipedia article” and “recount movie/book” prompts were promising, and we will continue to use more prompts like these in future work (selecting additional Wikipedia articles). We may further reduce the number of CPs collected by avoiding articles with spatially/visually descriptive topics. The very low CP/SR ratio of the “compare” and “personal intro/info” prompts was promising, and we will look for ways to encourage signers to elaborate further – to elicit longer stories when using these prompting strategies.

We plan on collecting/annotating approximately 200 ASL stories in total. Our experiences recording the first 58 stories have helped us to become more proficient at quickly and accurately collecting motion-capture data from signers, and we have developed new protocols for accurately and accessibly calibrating our equipment (Lu & Huenerfauth, 2009). We are also continuing to refine our annotation guide and training protocol for annotators to promote faster and more accurate annotation.

We are now beginning to analyze some collected 3D data to construct models of SRP establishment, spatial reference, and verb inflection. These models will be incorporated into ASL animation generation software we are developing to decide automatically: (1) when it should establish an SRP for an entity being discussed, (2) where it should place the SRP, and (3) how the signs later in the performance need to change based on SRP locations. In addition, we believe that our annotated ASL motion-capture corpus will be a valuable resource for future ASL linguistic researchers or computer scientists studying the synthesis of ASL animation or automatic recognition of ASL from human motion-data or video.

7. Acknowledgements

This research was supported by the U.S. National Science Foundation (award #0746556), Siemens (Go PLM Academic Grant), and Visage Technologies AB (free academic software license). Wesley Clarke, Kelsey Gallagher, Jonathan Lamberton, Amanda Krieger, and Aaron Pagan assisted with data collection/annotation.

8. References

- Brashear, H, Starner, T, Lukowicz, P, Junker, H. (2003). Using multiple sensors for mobile sign language recognition. *IEEE Intl Sym Wearable Computers*, p 45.
- Bungeroth, J, Stein, D, Dreuw, P, Zahedi, M, Ney, H. (2006). A German sign language corpus of the domain weather report. In C. Vettori (ed.) *2nd wkshp on represent. & processing of sign languages*, pp. 2000-3.
- Cormier, K. (2002). Grammaticalization of indexic signs: how American Sign Language expresses numerosity. Ph.D. Dissertation, University of Texas at Austin.
- Cox, S, Lincoln, M, Tryggvason, J, Nakisa, M, Wells, M, Tutt, M, Abbott, S. (2002). Tessa, a system to aid communication with deaf people. In *Proc. ACM Conference on Assistive Technologies*, pp. 205-212.
- Efthimiou, E., Fotinea, S.E. (2007). GSLC: creation and annotation of a Greek sign language corpus for HCI. In *LNCS 4554*, Heidelberg: Springer, pp. 657-666.
- Huenerfauth, M. (2009). Improving spatial reference in American Sign Language animation through data collection from native ASL signers. In *Proc. Universal Access in Human Computer Interaction*, pp 530-539.
- Huenerfauth, M, Hanson, VL. (2009). Sign language in the interface: access for deaf signers. In C Stephanidis (ed.) *Universal Access Handbook*. Mahwah: Erlbaum.
- Kennaway, J, Glauert, J, Zwitterlood, I. (2007). Providing signed content on Internet by synthesized animation. *ACM Transactions Computer-Human Interaction* 14(3), Article 15, pp. 1-29.
- Lenseigne, B, Dalle, P. (2005). A tool for sign language analysis through signing space representation. In *Proc. sign language linguistics and application of info. technology to sign languages, Milan, Italy*.
- Liddell, S. (2003). *Grammar gesture and meaning in American Sign Language*. UK: Cambridge Univ Press.
- Lu, P, Huenerfauth, M. (2009). Accessible motion-capture glove calibration protocol for recording sign language data from deaf subjects. In *Proc. ACM SIGACCESS Conference*, pp. 83-90.
- Meier, R. (1990). Person deixis in American sign language. In S. Fischer, P. Siple (eds.) *Theoretical issues in sign language research, Vol 1, Linguistics*. Chicago: University of Chicago Press, pp. 175-190.
- Neidle, C, Kegl, J, MacLaughlin, D, Bahan, B, Lee, R. (2000). *The syntax of ASL: functional categories and hierarchical structure*. Cambridge, MA: MIT Press.
- Padden, C. (1988). *Interaction of morphology and syntax in American Sign Language*. New York: Garland.
- Traxler, C. (2000). The Stanford achievement test, ninth edition: national norming and performance standards for deaf and hard-of-hearing students. *J Deaf Studies & Deaf Education* 5(4), pp. 337-348.
- VCom3D. 2010. Sign Smith Studio. <http://www.vcom3d.com/signsmith.php>. Accessed 11 March 2010.

The Icelandic sign language dictionary project: some theoretical issues

Nedelina Ivanova

Communication Centre for The Deaf and Hard of Hearing
Sudurlandsbraut 12, 108 Reykjavik, Iceland
E-mail: nedelina@visir.is

The Icelandic sign language dictionary project: some theoretical issues

This paper reports on the lexicographical description of the construction of an electronic dictionary for Icelandic Sign Language (ITM). The author reviews briefly some theoretical issues regarding the dictionary project: L1 Icelandic and L2 ITM and its potential users: the general public and the Deaf; the collection, evaluation and selection of signs; the lemmatizing process influenced by oral components on the semantic level and by manual features on the phonological level; the dictionary entry which is a sign demonstrated by a 'video clip'; access structures based on the specific phonological structure of SL, on the spoken language and picture themes with illustrations; the dictionary article where information about the nature of the signs is given and practical problems concerning the presentation of classifier predicates and the low reliability of hearing researcher moderating discussion sessions with Deaf informants is examined. The goal of the dictionary project is to collect the signs which are currently in use because there isn't a dictionary for ITM in order to (1) document the language and (2) be an instrument for researchers so that users can get practical avail of it and the dictionary will be of importance for getting legal recognition of ITM.

1. Introduction

There are approximately 300 Deaf users of Icelandic Sign Language (Íslenskt táknmál, ITM). The first dictionary of ITM was published in 1976 and was last edited in 1988. The ITM dictionary is a wordlist consisting of illustrations of the signs, sometimes specially invented for the list's purpose, presenting an Icelandic word or an inflected form of a common Icelandic verb and of loans from Swedish and Danish Sign Languages. In 2004 The Association of Parents and Benefit Society of Hard of Hearing children subsidized a compilation of signs which was published on the Internet under the name The sign bank. The novelty is that signs are shown by 'video clips'. Actual lexicographical work has not been done in this field in Iceland. These circumstances call for a compilation of an electronic dictionary of ITM based on linguistic principles and lexicographical methods.

The facts that dictionary compilation for SL is in general time-consuming, expensive and the limited number of potential users similarly to ITM make the work on a dictionary of ITM very difficult. The dictionary project for ITM has been more or less at a theoretical stage during the last two years, starting in 2008 with a M.A. thesis on lexicographical description for an electronic dictionary of ITM on the basis of linguistic principles (Ivanova, 2008) and in 2009 with a description of a lexical bilingual database for the dictionary compilation. At the same time in 2009 a list of 6441 signs was compiled by Deaf and hearing researchers at the Communication Centre for The Deaf and Hard of Hearing. Today in 2010 the project is on hold due to financial reasons.

However, the ITM dictionary project is the first incisive research on an ITM lexicon. The purpose of the ITM dictionary with its 4000 entries, when published, is to give answers concerning the basic forms, meanings and appropriate usage of the signs.

This paper reports only on the main lexicographical issues regarding the description for construction of the dictionary of ITM.

2. Theoretical issues

2.1 The dictionary and its potential users

The dictionary of ITM is bilingual, bidirectional and bifunctional (Svensén, 2004). The two languages are Icelandic or L1 and ITM or L2 where Icelandic is the mother tongue of the majority of potential users. The dictionary is L1→L2/L2→L1, both for hearing and Deaf people and both for perception and production of texts.

Hearing people can make use of the dictionary (1) to understand the meanings of signs and (2) to construct texts in ITM. Deaf users can make use of the dictionary (1) to understand the meanings of Icelandic words and (2) to produce texts in Icelandic by finding more equivalents to a sign, even though grammatical information for the equivalents is not given, at least not in the first edition.

Potential users of the dictionary include members of the general public interested in ITM; parents of Deaf children and their hearing friends, interpreters and hearing people teaching ITM, students in Sign Language studies, people who attend SL courses as well as the Deaf people themselves.

2.2 Sign's collection, evaluation and selection

The Deaf society is concentrated in the capital area and there aren't any regional variations of ITM, which made the collection process easier. 9616 signs were collected from (1) The ITM dictionary, (2) The sign bank, (3) various sign lists and (4) approximately 2 hours of video footage of conversations between Deaf people on different topics. Deaf researchers, divided in two groups by their age evaluated the 9616 signs according to five criteria: *current use by younger people*, *current use by older people*, *old sign, not in use* or *I do not understand the sign*. For the signs evaluated as *currently in use* the Deaf

researchers marked also the frequency of use according to their personal experience as *used by all* or *not used by all*. The two evaluations were compared for each sign and differences in them were discussed. The result is a list of 6441 signs including signs evaluated as *currently in use by younger and older people*, *used by all and not by all*, and *old signs*. It was decided to select 4000 signs evaluated as *currently in use by all younger and older people* for the first edition of the dictionary.

2.3 The lemma selection

The dictionary entry is a sign in its basic form demonstrated by a ‘video clip’ and an Icelandic gloss in the macrostructure of the dictionary. The basic form of the sign is „the simplest possible form of a lexeme which still identifies it uniquely and which still conveys what is regarded as its core or essential meaning.“ (Johnston & Schembri, 1999). It is not modified e.g. in plural or when inflected and it is the answer of the question: “What is the sign for ... ?”.

2.3.1. The lemma selection for lexical items with identical manual features

The lemmatization process is influenced by mouthings and mouth gestures as a lexicalized part of the lemma on the semantic level. The signs glossed in (1.a and b) differ only in mouthings, which imitate the Icelandic equivalent of the sign or a part of it. Mouthings are underlined in the examples:

- (1) a. SYSTKIN ‘brother(s) and sister(s)’
b. ALVEG SAMA ‘do not care’

The signs glossed in (2.a and b) differ only in mouthings:

- (2) a. BORÐA ‘eat’
b. MATUR ‘food’
c. NESTI ‘provisions’

The signs glossed in (3.a-c) differ in mouth patterns not related to Icelandic language and in imitations of sounds which do not constitute Icelandic word:

- (3) a. STRIÐA <ððððð> ‘tease’
b. HVERNIG <vo> ‘how’
c. AF HVERJU <hv> ‘why’

The signs glossed in (4.a and b) differ in mouth patterns not related to Icelandic language:

- (4) a. ÁST <munch> ‘love’
b. GÓÐ TILFINNING <neutral> ‘good feeling; good emotion’

The signs glossed in (5.a and b) differ in mouthing and mouth gesture:

- (5) a. DAGUR ‘day’
b. EKKERT <ððððð> ‘nothing’

The two signs in (1.a and b) are represented as two different lemmas in the dictionary, because their meanings are not connected (Berkov, 1996). The same principle applies to the signs in (3.a –c) and (5.a and b). The signs in (2 a.-c) are represented as one lemma with three different meanings, because of the relation of the meanings of these signs. The examples in (4.a and b) are treated in the same way. This decision was taken after numerous long discussions with Deaf researchers. Such

kind of distinction, where meanings are connected or not, could be very difficult to make and for some signs a compromise must be made at the expense of (1) more homonyms in the dictionary; (2) a more complex dictionary article for some signs and (3) a distinction between two or more dictionary entries which are treated as one sign by native speakers or vice versa, one dictionary entry and two or more signs.

2.3.2. The lemma selection for lexical items with identical meaning

The selection of lexical items with identical meaning is adopted from Troelsgård & Kristoffersen (2008). Signs are found to be synonyms on the basis of their phonological structure and are entered as two or more dictionary entries if they differ in two or more manual features: location, handshape, movement or orientation:
(6) a. ÞÚSUND ‘thousand’ S-handshape and movement down
b. ÞÚSUND ‘thousand’ T-handshape and movement forward

If signs differ only in one manual feature, they are treated as lemma and variant(s):

- (7) a. BRÁÐNA ‘melt down’ palm faces up
b. BRÁÐNA ‘melt down’ palm faces down

Frequency of use is determinant whether a sign is entered as lemma or as its variant(s).

2.4 Access structures

With the potential users in mind, access possibilities make the search for a sign easy and quick. The dictionary’s access structure requires (1) every sign’s phonological description and (2) grouping the signs in semantic fields. Searches are possible by four criteria based on the signs’ manual and non-manual features, Icelandic words or parts of words and illustrations.

Detailed phonological description for each sign will i.a. be the base for organizing the signs on the level on the macrostructure of the dictionary for ITM. A preliminary suggestion for a model for organizing the signs is based on the model for DSLD (Troelsgård & Kristoffersen, 2008) and on the description of phonological categories for the Sign Language of Netherlands (Crasborn, 2001; Van der Kooij, 2002).

2.4.1. Access by handshape

There are two possibilities for the user to access a sign by handshape. (1) If the handshape does not change during the production of the sign, the user can choose a handshape or variant of the handshape for the strong and/or for the weak hand from a set with handshapes (e.g. *Suvi* and The Danish Sign Language Dictionary, abbreviated as DSLD). In two-handed signs the chosen handshape may be the same as for the strong hand or not. The user gets a sign or list of signs, both one-handed signs and two-handed signs, which have the chosen handshape or handshapes. (2) If the handshape does change, the user can choose a handshape or variant of handshape for the strong hand for the beginning of the production of the sign

from a set with handshapes and then he can choose another handshape at the end of production of the sign from a popup window with suggestions of possible handshape combinations. Those suggestions are based solely on the phonological information about the signs included in the dictionary. It is not expected from the user to analyze the signs, but to find the sign he might be looking for as quickly as possible. The user gets a sign or a list of signs which have the chosen handshape for the beginning and the chosen handshape at the end of the production of the sign.

An informal research at the Communication Centre for The Deaf and Hard of Hearing has shown that there are about 40 handshapes in ITM, but this issue still needs to be researched.

2.4.2. Access via location

Here the user can choose again between two possibilities to access a sign. (1) He chooses a location from a set of pictures for different locations (e.g. *Suvi* and *DSL*). (2) If the location does change during the sign's production a popup window opens with suggestions of possible end location. The user combines both locations. He gets a sign or a list of signs which have the chosen location or combination of locations.

It is also possible to combine a handshape or handshapes with location and search for sign(s) which have the chosen combination of handshape(s) and location.

At this stage 25 locations are defined. However, more research needs to be conducted.

2.4.3. Access by mouth gestures with no relation to Icelandic language

The user gets a list of all mouth gestures with no relation to Icelandic language which are to be found in the dictionary. He chooses a mouth gesture. He gets an exhaustive list of all signs that have the chosen mouth gesture.

2.4.4. Access by mouth gestures which are imitations of sounds that do not constitute Icelandic words

The search principle is the same as in 2.4.3.

Research on mouth gestures has not been done yet so it is not possible to say how many they are.

2.4.5. Access by an Icelandic word

The user may search for a sign by typing in an Icelandic word or a phrase. The search box displays a list of suggestions to assist the user in finding a word or phrase. The search is in the equivalents, in explanations and glosses for the examples. The user gets a list which includes all the dictionary entries which match the typed word or phrase. Icelandic equivalents which are nouns are given in nominative singular; adjectives are in nominative singular masculine and verbs are in infinitive, i.e. the equivalent's form is not inflected for case, number, gender and time. The same principle applies also to glosses in Icelandic.

2.4.6. Access via picture themes with illustrations

The idea is adopted from the LEXIN dictionaries¹. Signs are grouped in picture themes for concrete phenomena on the basis of collective interrelation to the topic in question. An illustration of the phenomenon is to be found in the picture theme it belongs to. Access to the dictionary entry is through the illustrations. After choosing a picture theme the user gets a collection of smaller illustrations which characterize that theme. The user chooses an illustration by clicking on it. The equivalent sign opens in a popup window. The sign is demonstrated by a 'video clip' and an Icelandic gloss. The Icelandic gloss is linked to the relevant dictionary article in case the user would like to read more about the lemma. The use of such kind of access to the dictionary leads to avoidance of the written Icelandic word as an entry to a sign. This access can be used e.g. by parents of Deaf children, who only want to see the sign and not the dictionary article, by Deaf children and children of Deaf adults in order to increase their vocabulary, and by Deaf foreigners who do not know Icelandic.

2.5 The dictionary article

In the dictionary article phonological information is given with pictures which show two manual features of the sign: handshape and location (as in *Suvi* and *DSL*).

Mouthings are shown by underlining that part of the Icelandic equivalent which is "pronounced". Mouth gestures are described and shown in $\langle \rangle$. A sign's meaning is given by Icelandic equivalent(s) or explanation(s). A sign's modification for plural is shown by a link to the correspondent part in the explanatory grammar chapter in the dictionary. A sign's modification for subject-object verb agreement is illustrated through example. The example in the dictionary article consists of a 'video clip', a gloss of the example in Icelandic and translation in Icelandic. Variants of the sign are marked and are shown with a 'video clip'. Links in the dictionary article lead to homonyms, synonyms and picture theme when applicable. Information on a lemma's area of use, limitations of use and shades of meaning are also given when applicable.

2.6 Practical problems

2.6.1. Classifier predicates in the dictionary article

Being part of the productive lexicon the classifier predicates are not given the status of lemmas in the dictionary. They are shown in the dictionary article in form of examples of the use of the dictionary entry. With potential users in mind and their knowledge or lack of knowledge of sign language grammar and terminology it is hard to find a right way to gloss the meaning of classifier predicates. Two ways are considered possible: (1) to write the word 'proform' in the gloss (e.g. as in

¹ The LEXIN dictionaries are web-based dictionaries for immigrants in Norway, Sweden and Denmark. In those dictionaries picture themes with illustrations are also used to access lemmas. The Icelandic LEXIN project is on hold as of 2010.

DSLID) without any explanations and a translation in Icelandic presenting the meaning of the classifier predicate or (2) to gloss the classifier predicate with small letters as simple as possible and concisely enough to present the meaning. The meaning of the classifier predicate is given as a combination of the translation in Icelandic and the video clip. For the Icelandic dictionary the second possibility was chosen even though this approach is known to be time-consuming and quite challenging.

2.6.2. Low reliability of hearing researcher moderating discussion sessions with Deaf informants

In trying to extract the potential meaning(s) of a sign and its use two discussion sessions with Deaf informants (the same researchers who evaluated the signs and are familiar with the project) and moderated by hearing researcher were held. A sign (or a root?) in its basic form, but without mouthings and mouth gestures, was presented. Deaf informants were asked (1) to suggest which sign(s) might have the concrete manual structure, (2) to accompany the sign with proper mouthings and/or mouth gestures and (3) to use the sign(s) in context. In this preliminary research was noticed that it had would be better if Deaf researcher moderated the sessions for two reasons: (1) The sign language used in these two sessions by Deaf informants with the hearing researcher differed in structure from the sign language Deaf people used between themselves. It was strongly influenced by Icelandic grammar and the meanings of the words in Icelandic. (2) Deaf informants tried to give answers and examples of what they thought the hearing researcher was looking for instead of using the signs being researched in context in ITM.

3. Conclusion

As shown in this paper, the project for a dictionary of ITM is at a planning stage, i.e. it is based mostly on theory and very little on practice. It is conceivable that some of the issues described in this paper are really hard to achieve, more time-consuming than was thought in advance and changes would be necessary. The dictionary project for ITM does not aim to be a novelty in the field of SL lexicography because ideas from dictionaries of other sign languages have been adopted, but the dictionary project is novelty for ITM being the very first lexicographical project and therefore of importance for (1) documentation and basic research of ITM and (2) getting legal recognition of the language.

Acknowledgements

The work on the project reported here was supported in part by a grant from the Fund for Non-fiction Writers and by the Communication Centre for The Deaf and Hard of Hearing. Thanks to Jette Kristoffersen for constructive discussions.

References

- Berkov, V. P. (1996). *Dvujazytsjnaja leksikografija: utsjebnik*. Sankt-Peterburg: Izd-vo S.-Peterburgskogo universiteta.
- Crasborn, O. A. (2001). *Phonetic implementation of phonological categories in sign language of the Netherlands*. Published Ph.D. Dissertation. LOT Dissertation Series 48. LOT, Utrecht, The Netherlands. Online publication: www.lotpublications.nl/publish/issues/Crasborn/index.html
- DSLID = *Ordbog over Dansk tegnsprog*. (2008). Kristoffersen, J. (ed.). Copenhagen: Centre for Sign Language and Sign Supported Communication – KC. Online publication: www.tegnsprog.dk
- Ivanova, N. (2008). Proposition for new ITM dictionary on linguistics principles. Unpublished M.A. thesis, University of Iceland, Reykjavik.
- Johnston, T., Schembri, A. (1999). On Defining Lexeme in a Sign Language. *Sign Language and Linguistics* 2 (2), pp. 115--185.
- LEXIN = *Lexin ordbøker for innvandrere*. (2007). Bjørneset, T. (project leader). Bergen: Utdanningsdirektoratet, Uni Digital. Online publications:
Norway: <http://decentius.hit.uib.no/lexin.html>
Sweden: <http://www-lexikon.nada.kth.se/skolverket/>
Denmark: <http://lexin.emu.dk/>
Iceland: http://www.lexis.hi.is/lexin_ny.html
- Suvi = *Suomalaisen viittmakielen verkkosanakirja*. (2003). Helsinki: Finnish Association of the Deaf. Online publication: suvi.viittomat.net
- Svensén, B. (2004). *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. Andra, omarbetade och utökade upplagan. Stockholm: Norstedts Akademiska Förlag.
- The sign bank = *Táknabankinn*. (2004). Samstarfsverkefni Foreldra – og styrktarfélags heyrnardaufra, Félags heyrnarlausra og Samskiptamiðstöðvar. Reykjavik. Online publication: www.taknmal.is
- The ITM dictionary = *Táknmálsorðabók*. (1987). Reykjavik: Félag heyrnarlausra.
- Troelsgård, T., Kristoffersen, J. (2008). An electronic dictionary of Danish Sign Language. In Müller de Quadros . R. (ed.), *Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th Theoretical Issues in Sign Language Research conference*. Florianopolis, Brazil, December 2006, pp. 652--662. Online publication: www.editora-arara-azul.com.br/ebooks/catalogo/abertura.pdf
- Van der Kooij, E. (2002). *Phonological Categories in Sign Language of the Netherlands The Role of Phonetic Implementation and Iconicity*. Published Ph.D. Dissertation. LOT Dissertation Series 55. LOT, Utrecht, The Netherlands. Online publication: www.lotpublications.nl/index3.html

A Comparison of Two Linguistic Sign Identification Methods

Tommi Jantunen

University of Jyväskylä, Department of Languages
P.O. Box 35 (F), FI-40014 University of Jyväskylä, Finland
E-mail: tommi.j.jantunen@jyu.fi

Abstract

This paper employs two linguistic sign identification methods – a manual one focusing on the dominant hand and a nonmanual one focusing on the mouth – and compares the kinds of sequences they classify as signs from a video containing continuous signing. The study is motivated by two projects, of which one investigates the ontological nature of the sign and the other aims to develop an automatic sign recognition tool. In the study, both methods were able to associate all the free semantic-functional elements in the data with signs. However, in the nonmanual method the overall number of identified signs was lower because the stretching of the mouth movement of the semantic element over the following pointing meant that the combinations of semantic elements and pointings were counted as single signs. Moreover, signs identified by the nonmanual method were longer than those identified by the manual method. The results from the nonmanual method agree with the claim that phrase internal sequences of semantic elements and pointings are lexical head plus clitic combinations. Consequently, it is suggested that pointings in such contexts do not need to be independently detected by the automatic sign recognition tool.

1. Introduction

This paper presents a study that employed two different linguistic sign identification methods and compared the kinds of sequences they identified as signs, especially in terms of the relative length of the sequences, from a video signal containing continuous signing. The first method focused on the dominant hand and is referred to in this paper as the *manual method*. The second method focused on the mouth and is referred to as the *nonmanual method*. Both methods were applied to a small set of data extracted from the Basic Dictionary of Finnish Sign Language (FinSL) signed example text corpus, publicly available through *Suvi* (<http://suvi.viittomat.net>).

The study is motivated by two projects currently underway in research into FinSL. The first project is a linguistic one, aiming to test empirically certain ontological assumptions concerning three linguistic units – the sign, the syllable, and the sentence – in signed language research (see <http://users.jyu.fi/~tojantun/3BatS>). In the project, the notion of the sign is taken as the reference point to which all other notions are proportioned. Consequently, in order to carry out the project successfully, the empirical nature of the sign must first be explored. Comparing the results of two different linguistic sign identification methods contributes to the completion of this particular task.

The second project is a technological one, aiming to develop content-based video analysis methods and an automatic sign recognition tool for FinSL (Koskela *et al.*, 2008). As a starting point it has been assumed that the detection of signs from a video requires the use of several technologies, such as a dominant hand motion detector and a mouth movement or position detector. In order to successfully develop these technologies it is necessary to first describe and evaluate the kinds of se-

quences that can be expected to be classified as signs by observing the dominant hand and the mouth independently; and by a human linguist.

The two projects are interconnected in that the first project feeds the second with linguistic substance while the second project provides technological analysis tools for the first. So far, this cooperation has been successful as we have already been able to develop a method that enables a sign language researcher to graphically represent and semi-automatically analyze signed language motion from digital video material containing natural signing (Jantunen *et al.*, forthcoming). This method, in combination with the PicSOM retrieval system framework for content-based analysis of multimedia data (<http://www.cis.hut.fi/picsom/>), will be investigated further to develop a dominant hand motion detector and an automatic sign recognition tool for FinSL. The PicSOM system will also be adapted to recognise the shapes of mouth movements and positions (Koskela *et al.*, 2008).

2. The sign identification methods

The creation of signed language corpora in different countries has made it necessary to spell out the linguistic methods used in identifying signs from a video. In determining the beginnings and ends of signs most methods take the dominant hand, i.e. the most salient articulator in signed language, as the reference point (although they usually describe the dominant and nondominant hand on separate tiers; e.g. Crasborn & Zwitterlood, 2008; Johnston, 2009). The dominant hand is the reference point also in the manual method used in the present study. The second method, on the other hand, relies on observing the movements and positions of the mouth (i.e. mouthings and mouth gestures). The motivation for this nonmanual method stems from the fact that FinSL signs are accompanied with a mouth movement or position of some sort and that these either differentiate between or

specify the meanings of FinSL signs (Rainò, 2001). It is therefore argued that signs are linguistically identifiable through observing the actions of the mouth.

In the manual method, the beginnings and ends of signs are determined on the basis of changes occurring in path and local movements produced by the dominant hand. The beginning of a sign is taken to correspond to the video frame that immediately precedes the frame in which the dominant hand first shows movement away from the initial location of the sign. If the sign includes only a local movement, the beginning of a sign corresponds to the frame that immediately precedes the frame in which the initial handshape or orientation of the dominant hand first starts to change. A sign is taken to end at the frame in which the path movement of the dominant hand has reached its end or in which the dominant hand still holds a posture or a hand configuration of the sign.

In the nonmanual method, a sign is taken to start from the frame that is associated with the moment the mouth has acquired the initial position for the mouthing or mouth gesture to be recognisable. A sign ends at the frame that corresponds to the completion moment of the mouthing or mouth gesture. Should the activity of the mouth be unobservable (e.g. due to occlusion by the hand), the manual method will be used for the beginning and/or end of that particular sign.

The temporal start and end moments of signs indicated by the two methods are not assumed to be absolute. The relative nature of the beginnings and ends of signs is emphasised especially by the identification of two-handed signs in the manual method. In two-handed signs, both hands may move or hold a posture independently, in which case the beginning or end moments of these signs would be best determined by analysing both hands separately. However, in this paper two-handed signs are treated only in terms of their dominant hand.

3. The data

The data for the present study was extracted from the Basic Dictionary of FinSL signed example text corpus (the BDFinSL corpus; cf. *Suvi*). Altogether the corpus consists of roughly 5000 video clips (25 fps) each identifiable as one signed sentence or minitext. The sentences/minitexts were prepared by native deaf FinSL signers with the objective of creating a context as natural as possible for the lexemes presented in the dictionary. The corpus is assumed to represent the standard everyday variety of FinSL although it is likely to put slightly more emphasis on the variety used in southern Finland.

From the roughly 5000 video clips of the BDFinSL corpus data, a smaller set of 60 clips was first extracted by systematically selecting the second clip of every 20th lexical entry in the dictionary; this set was collected for use later in another study. After this, five clips were ex-

tracted from the set of 60 clips by using simple random sampling. These clips turned out to be examples 500/2, 660/2, 800/2, 860/2, and 1120/2 of the BDFinSL corpus (the number of the lexical entry in the dictionary/the number of the example clip in each entry) and they formed the data for the present study. The clips were opened in Apple's QuickTime Pro application (version 7.6.4) on a Macintosh computer and subjected to the manual and nonmanual sign identification methods described in Section 2. The start and end frames of signs were identified by observing the (absolute) frame number indicator of the QuickTime Pro application.

4. The results of the comparison

The results of the study are displayed in Tables 1–5 for examples 500/2, 660/2, 800/2, 860/2, and 1120/2, respectively. The left hand column in each table contains a short characterisation of all the free semantic and functional elements (cf. non-bound sequential morphemes and gestures) present in each example, identified prior to the application of the two methods. Each characterisation describes either the rough basic meaning of the element (e.g. 'girl') or the function of the element (e.g. pointing). The epithets occurring after pointings specify the referent of the pointing (e.g. 'me') or the relative direction of the pointing (e.g. left); an additional epithet "-go" in pointings indicates that the pointing has a verbal reading. The middle and right hand columns display first the interval of frames that contain the sign as identified by the manual and nonmanual method respectively. Each interval marker is followed by a number in parenthesis that indicates the length of the sign in terms of frames.

Element	Signs M	Signs NM
'girl'	37-40 (4)	
pointing-left	45-47 (3)	36-49 (14)
'party'	52-55 (4)	
pointing-left-go	59-64 (6)	51-66 (16)
'cannot'	70-77 (8)	
pointing-left-go	79-82 (4)	70-82 (13)
'because'	90-95 (6)	86-95 (10)
pointing-left	98-99 (2)	97-99 (3)
'agree'	103-106 (4)	101-109 (9)
'already'	113-121 (9)	112-122 (11)
'children'	126-131 (6)	124-132 (9)
'care'	135-138 (4)	
pointing-right-go	146-150 (5)	134-151 (18)

Table 1: The results in frames of the manual (M) and nonmanual (NM) method for example 500/2.

Table 1 displays the results for example 500/2 of the BDFinSL corpus. The manual method identified all the 13 free semantic and functional elements of the example as signs. The number of signs identified by the nonmanual method was 9. The nonmanual method did not

leave out any semantic or functional elements but it counted the phrase-internal sequences of semantic elements and pointings as single signs. This was due to the stretching of the mouth movements of semantic elements over pointings (see e.g. Rainò 2001): for example, the Finnish mouthing [eei.vo] originating from the Finnish words *ei voi* 'can not' was stretched over the sequence 'cannot'+pointing-left-go in such a way that the first syllable of the mouthing was associated with the element 'cannot' and the second syllable with the element pointing-left-go. The mean length of a sign identified by the manual method was 5 frames (SD=2) and the mean length of a sign identified by the nonmanual method was 11.4 frames (SD=4.4); if the signs consisting of a semantic element and a following pointing are left out of the count, the mean length of a sign identified by the nonmanual method drops to 8.4 frames (SD=3.1).

Element	Signs M	Signs NM
'my own'	37-40 (4)	35-41 (7)
'father'	43-49 (7)	43-57 (15)
pointing-right	52-56 (5)	
'no'	63-67 (5)	61-67 (7)
'my own'	70-73 (4)	69-74 (6)
'father'	80-85 (6)	77-84 (8)
'half'	87-96 (10)	86-100 (14)

Table 2: The results in frames of the manual (M) and nonmanual (NM) method for example 660/2.

Table 2 shows the results for example 660/2 of the BDFinSL corpus. Here again the manual method identified all the 7 semantic and functional elements of the example as single signs whereas the number of signs identified by the nonmanual method was 6. In the nonmanual method, the phrase-internal sequence of the semantic element 'father' and the following pointing was counted as a single sign, due to the stretching of the mouthing over the pointing. The mean length of a sign identified by the manual method was 5.9 frames (SD=2.1) whereas the mean length of a sign identified by the nonmanual method was 9 frames (SD=3.8); if the one sign consisting of two elements is left out of the count, the mean length of a sign identified by the nonmanual method in this example drops to 8 frames (SD=3).

Table 3 displays the results for example 800/2 of the BDFinSL corpus. The number of signs identified by the manual method is 8, corresponding to the number of free semantic and functional elements in the example. The number of signs identified by the nonmanual method is 7 because the final combination of a semantic element ('lose opportunity') and a pointing are counted as one sign. The mean length of a sign identified by the manual method was 4.9 frames (SD=2) whereas the mean length of a sign identified by the nonmanual method was 9.1 frames (SD=6.5); if the one sign consisting of two semantic-functional elements is left out of the count, the

mean length of a sign identified by the nonmanual method in this example drops to 7 frames (SD=3.5).

Element	Signs M	Signs NM
'talk'	43-45 (3)	42-45 (4)
'should have'	47-51 (5)	47-51 (5)
'no'	56-60 (5)	54-61 (8)
'have to'	74-77 (4)	71-78 (8)
'underwrite'	82-90 (9)	81-93 (13)
pointing-me	93-95 (3)	92-95 (4)
'lose opportunity'	102-107 (6)	102-123 (22)
pointing-me	113-116 (4)	

Table 3: The results in frames of the manual (M) and nonmanual (NM) method for example 800/2.

Table 4 presents the results for example 860/2 of the BDFinSL corpus. The number of signs identified by the manual method was 5 (i.e. all the free semantic and functional elements) and the number of signs identified by the nonmanual method was 3. In the nonmanual method, the sequence of the first two semantic-functional elements of the example ('believe' and the following pointing) as well as the sequence of the two final elements ('no' and the following pointing) were counted as single signs due to the spreading of the mouth movement and position respectively. The mean length of a sign identified by the manual method was 4.8 frames (SD=1.5) whereas the mean length of a sign identified by the nonmanual method was 17 frames (SD=10.4) (the length of the one sign not including two elements was 5 frames).

Element	Signs M	Signs NM
'believe'	38-43 (5)	30-51 (22)
pointing-you	47-51 (5)	
'come along'	57-59 (3)	56-60 (5)
'no'	65-71 (7)	63-86 (24)
pointing-you	78-81 (4)	

Table 4: The results in frames of the manual (M) and nonmanual (NM) method for example 860/2.

Finally, Table 5 displays the results for example 1120/2. The number of signs identified by the manual method was 5 and the number of signs identified by the nonmanual method was 4 (cf. 'obscene'+pointing-left). The mean length of a sign identified by the manual method was 9 frames (SD=3.4) whereas the mean length of a sign identified by the nonmanual method was 18.3 frames (SD=10.2); the mean length of a sign identified by the nonmanual method without the one two-element sequence drops to 13.7 frames (SD=5.5).

To conclude, both methods were able to identify all the free semantic and functional elements in the examples. However, the methods produced different results with

Element	Signs M	Signs NM
'who'	33-39 (7)	31-40 (10)
'draw'	45-53 (9)	44-54 (11)
'painting'	62-75 (14)	57-76 (20)
'obscene'	86-95 (10)	
pointing-left	102-106 (5)	80-111 (32)

Table 5: The results in frames of the manual (M) and nonmanual (NM) method for example 1120/2.

respect to the element-sign ratio. To be more precise, the overall number of signs identified by the nonmanual method was lower because the stretching of the mouth movements and positions of the semantic elements over the pointings meant that the sequences of semantic elements and pointings were identified as single signs. Furthermore, signs identified by the manual method were relatively short in terms of frame count whereas signs identified by the nonmanual method were long: the total mean length of a sign identified by the manual method was 5.9 frames (SD=1.8) whereas the total mean length of a sign identified by the nonmanual method was 13 frames (SD=4.4); the total mean length of a sign identified by the nonmanual method without the two-element combinations was 8.4 frames (SD=3.2). When compared to the signs identified by the manual method, the signs identified by the nonmanual method typically contained, with the exception of example initial and final signs (see Tables 1–5), one to two additional frames both at the beginning and at the end of each identified sequence.

5. Discussion and conclusion

In general, the results agree with the assumption (see Section 2) that both the beginnings and ends of signs and, consequently, also the concept of (a linear) sign are indeed largely relative notions: for example, the fact that the total mean length of a sign can be either 5.9 or 13 frames (or 8.4 frames) demonstrates that what counts as a sign depends, among other things, on the sign identification method. This conclusion has been further strengthened during discussions with native FinSL signers. When asked to judge the sign-likeness of the signs identified by the two methods, the signers have accepted both types of sequences as signs. Interestingly, however, signs identified by the nonmanual method have been judged to be "more complete" because of the more visible mouthing / mouth gesture. Obviously, the existence of pointings in double element signs has been noticed but this has not led to the rejection of the sign-likeness of the sequences. This is additional evidence for the claim that pointings in these contexts function as grammatical clitic-elements attached to lexical heads (e.g. Zeshan, 2002; Jantunen *et al.*, forthcoming), not as pure signs.

The fact that both linguistic methods were able to associate all the free semantic and functional elements in the data with signs seems at first to suggest that the development of the automatic sign recognition tool for FinSL

could be based independently on either of the two methods; this is contrary to the initial assumption of the technological project outlined in Section 1. However, a closer look at the results indicates that, for the successful detection of signs from the video, a technology combining both methods is important. For example, the identification of durationally short signs (e.g. ≤ 5 frames) might not be possible if the recognition technology is based only on the manual method. On the other hand, a sign recognition technology based on only the nonmanual method cannot identify individual pointings closely following semantic elements. Interestingly, however, the present data regarding the clitic (i.e. non-sign) characteristics of pointings suggests that pointings in these contexts do not perhaps need to be separately detected by the automatic sign recognition tool at all. This possibility must be taken more seriously into account in the development of the tool.

6. Acknowledgements

I wish to thank Tuija Wainio for the valuable discussions I had with her in preparing the present paper. The financial support of the Academy of Finland is gratefully acknowledged.

7. References

- Crasborn, O. & Zwitserlood, I. (2008). Annotation of the video data in the "Corpus NGT". Dept. of Linguistics & Centre for Language Studies, Radboud University Nijmegen, The Netherlands. Online publ. <http://hdl.handle.net/1839/00-0000-0000-000A-3F63-4>
- Jantunen, T., Koskela, M., Laaksonen, J. & Rainò, P. (forthcoming). Towards automated visualization and analysis of signed language motion: Method and linguistic issues. To appear in the *Proceedings of 5th International Conference on Speech Prosody*, Chicago, Ill. (USA), May 2010.
- Johnston, T. (2009). Guidelines for annotation of the video data in the Auslan Corpus. Dept. of Linguistics, Macquarie University, Sydney, Australia. Online publ. http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf
- Koskela, M., Laaksonen, J., Jantunen, T., Takkinen, R., Rainò, P. & Raika, A. (2008). Content-based video analysis and access for FinSL - a multidisciplinary research project. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood (Eds.), *Construction and exploitation of sign language corpora*. Paris: ELRA, pp. 101–104.
- Rainò, P. (2001). Mouthings and mouth gestures in Finnish Sign Language. In P. Boyes Braem & R. Sutton-Spence (Eds.), *The hands are the head of the mouth: The mouth as articulator in sign languages*. Hamburg: SIGNUM-Press, pp. 41–50.
- Zeshan, U. (2002). Towards a notion of 'word' in sign languages. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *Word. A cross-linguistic typology*. Cambridge: Cambridge University Press, pp. 153–179.

Requirements For A Signing Avatar

Vince Jennings, Ralph Elliott, Richard Kennaway, John Glauert

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

{V.Jennings, R.Elliott, R. Kennaway, J.Glauert}@uea.ac.uk

Abstract

We present the technical specification for an avatar that is compliant with Animgen, the synthetic signing engine used at the University of East Anglia for generating deaf signing animations. The specification will include both the basic definition required for any standard animating avatar, and the additional parameters that Animgen requires to generate signing. Avatars compatible with Animgen are created using the ARPToolkit, an application developed at UEA that has a plug-in architecture for tools that are used for rigging an avatar mesh for animation. The toolkit also generates the additional data needed by Animgen for each avatar.

1. Introduction

For any avatar to be animated there is a standard set of requirements that must be met in the avatar file, which must include a mesh, a skeleton, a texture, and, if facial animation is also required, a set of morph targets. The mesh represents the visible shape of the avatar, and, together with the texture, defines the avatar's appearance. The skeleton, a mathematical construct in software which is not visible, has its bones linked to the vertices in the mesh, so that changing the rotation of any bone in the skeleton results in the movement of the mesh vertices linked to it. The morph targets, also referred to as blend shapes, each represent a deformation of the static mesh to both the area around the mouth and jaw for speech synchronisation, and to the cheeks, eyelids, eyebrows, and forehead for facial expressions.

JASigning is a synthetic animation system for deaf signing, written in Java, that has been developed at UEA, taking as input avatar-independent Gestural SiGML (Signing Gesture Markup Language) (Elliott *et al*, 2004, 2007) and producing as output motion data for any avatar. SiGML is an XML form of HamNoSys (Hamburg Notation System) (Prillwitz *et al*, 1989; Hanke, 2004) that is used by Animgen (Kennaway, Glauert, Zwitserlood, 2007) to generate signing animation. To achieve this JASigning requires additional information that cannot be obtained from the standard information above, and must be provided in separate files. To demonstrate the need for the extra data an example would be where a sign requires that the tip of the index finger on the right hand touches the tip of the nose. These locations cannot be obtained from the standard specification, but are provided in the extra files.

The ARPToolkit [ARP] was developed at UEA to provide a unified application for creating avatars that not only met the standard requirements for animation but also have the additional data needed for deaf signing. Additionally, the tools developed in the toolkit were designed to automate some of the tasks of avatar rigging, and to provide simple interfaces for some of the more complex tasks, such as morph target creation, making the toolkit accessible to users who lack the

technical skills needed for the majority of commercial software that would otherwise have to be employed.

For the purposes of the JASigning software, each ARP signing avatar is effectively defined by a set of four avatar definition files.

The first of these contains binary data, the other three are XML:

- Main Avatar Definition
- ASD, Avatar Standard Description
- Animgen Configuration Data
- Nonmanuals

2. Main avatar definition file

The main avatar definition file, **avatardef.arp**, contains only the data needed for an avatar to perform standard animations, and has none of the extra data that Animgen needs for the generation of deaf signing. Its major components are:

2.1 Vertex List

A list of vertices that represents the mesh defining the shape of the avatar, with texture coordinates and vertex normals for each. Each vertex and normal is defined relative to its linked bone(s), with the bone initially aligned along the X axis. Meshes for the eyes, teeth, and tongue must be present, but not contiguous with the rest of the mesh. The hair and ears can also be modelled separately from the main mesh, but all other parts of the avatar mesh must be a single contiguous mesh. To allow realtime animation (at 25fps or more) on an average specification machine the vertex count of the mesh should not exceed 10,000.

2.2 Texture Map

The texture map, which may optionally be held in a separate file or embedded in the file in a standard format such as PNG, defines the appearance of the avatar. All texture should be contained in a single file. For good quality a minimum size of 1024 X 1024 pixels is suggested.

2.3 Skeleton

The skeleton structure fits within the mesh, and includes bones for animating the eyes, which must be child nodes of the head bone. It must include all bone names used by Animgen (see *asd.xml* below), but can include additional bones (e.g. metacarpals), although these will be ignored by Animgen. Bone names are all 4 character (4cc) codes.

The bone hierarchy, as specified in the *asd.xml* file, must be adhered to, but is compatible with other standard hierarchies such as H-Anim and BVH. Translations and rotations for each bone are in the parent's coordinate space, with the transform for each bone being multiplied by its parent

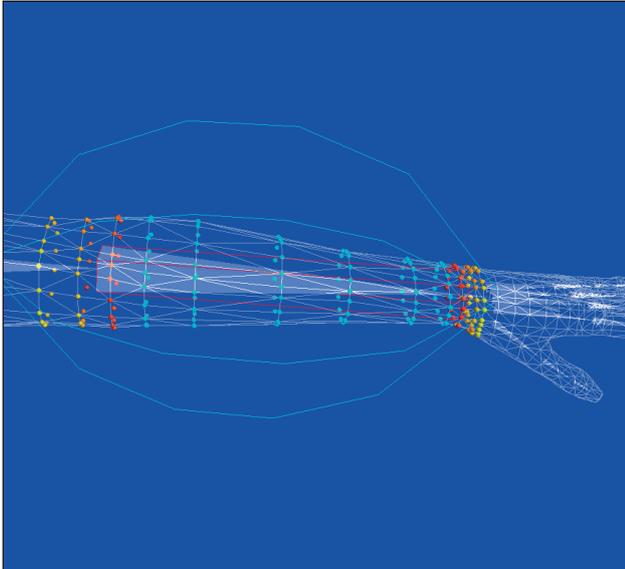


Figure 1. Mesh to skeleton attachment

bone's transform. Zero length bones may be included on all leaf nodes in the bone hierarchy if desired, again with unique 4 character names. These are sometimes required in other animation applications, and will also be ignored by Animgen. The ARPToolkit has tools for creating skeletons and for adjusting them to fit the avatar mesh.

2.4 Mesh-to-Skeleton Attachment Data

This data is a list of links between vertices in the mesh and the bone(s) that will animate them, with a weight for the influence of each bone. A maximum of 4 links per vertex is permitted, with a preferred maximum of 3. The weights of all links to a vertex must sum to 1.0. This follows standard industry practice for this type of data as more than 4 links to a vertex makes weight calculations very complex. Vertex weights are calculated in the ARPToolkit during construction of the skeleton, and their weights subsequently altered to produce good deformation at the joints by 'painting' the weights for each vertex using the mouse.

Figure 1 shows the linkage between the vertices of the arm and the bone. The envelope determines which vertices are assigned to the bone, and the colour of each vertex shows the weight (between 0 and 1) that this bone applies to the transformation for this vertex. These are typically 1 across the centre of the bone, reducing at the joints where adjacent bones also apply their weights.

2.5 Morph Targets

The list of morph targets contained in a standard non-signing avatar file would consist of the visemes necessary for lip synchronisation to speech, and for facial expressions. Each morph target represents a deformation of the mesh to produce facial animation. A morph target includes a list of indices for vertices in the mesh, a deformation vector for the full displacement of the vertex (1.0), and a normal for the fully displaced vertex. Negative amounts for morphs are not supported, e.g. for moving eyebrows down instead of up, so morphs for all movement directions must be provided.

For a deaf signing avatar additional morph targets are

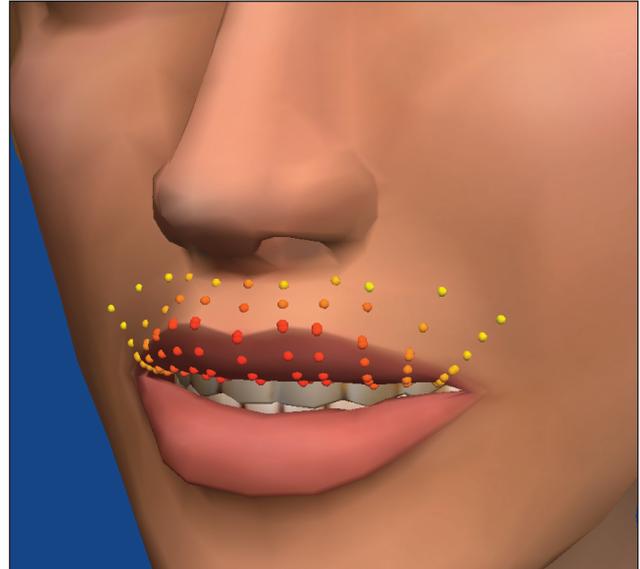


Figure 2. Upper lip morph target

needed, particularly for the cheeks and the tongue, which are used for a wide range of facial gestures.

Morph names are arbitrary, and can be matched to those used by Animgen, the synthetic signing engine used in JASigning, by editing the avatar's nonmanuals.xml file.

Morph targets are created in the ARPToolkit, where a library of primitive morphs are first defined for movements of the jaw, lips, tongue, cheeks, nose, eyebrows, and eyelids. Selections of these primitives are then combined to produce the morphs for phonemes and signing mouth gestures and uploaded into the avatar.

Figure 2 shows the vertex selection for the morph primitive for the upper lip, with the colour coding indicating the weighting of the transform that moves the vertices vertically, with red indicating a heavier weight falling off to yellow for a low weight.

3. The ASD File Format

The purpose of the Avatar Standard Description (ASD) file, **asd.xml**, is to define all the avatar-related data needed by Animgen. It defines the skeleton, with the bone names and hierarchy used by Animgen, in a reference pose that enables Animgen to establish the correct rotation axis for elbows and thumbs.

The ASD file also defines a set of approximately 380 feature points on the surface of the mesh of the upper body, arms, hands, and head, which Animgen may use as reference points when it needs to determine locations in signing space. On the arms and hands these points are defined on 2 axes at each joint and again midway between each joint. Each of these points is assigned a unique identity code recognised by Animgen.

To simplify the task of defining the feature points, which would otherwise have to be defined individually by hand, tools have been developed in the ARPToolkit to carry out ray tracing from the bones of the skeleton to intersect with the mesh at the desired locations. This process is automatic for the upper body, arms, and hands, with a secondary as-



Figure 3. Feature points with reference pose

sisted manual process for the head locations.

In the section of an file `asd.xml` shown below, the skeleton hierarchy is shown by the joint relationship “ROOT”, “SPI1”, “SPI2”, etc, with feature points being listed under their “owner” bones. Positions of each point are relative to their “owner”. For example T-LS is “torso front at left shoulder”, and its “owner” is “ROOT”.

```
<?xml version="1.0" standalone="yes"?>
<avatarStaticData version="1.0">
  <avatar name="arp-anna" version="1.0">
    <skeleton scale="0.04445039">
      <joint name="ROOT" position="0.000000 0.000000 0.000000" rotation="0.000000 0.000000 0.707330 0.706883">
        <feature name="T-LS" position="9.430 -4.333 1.636" />
        <feature name="T-CS" position="9.433 0.000 1.932" />
        <feature name="T-RS" position="9.436 4.333 1.638" />
        <feature name="T-LC" position="7.265 -2.168 3.171" />
        <feature name="T-CC" position="7.266 0.000 3.284" />
        <feature name="T-RC" position="7.268 2.165 3.175" />
        <feature name="T-LA" position="4.269 -2.166 2.485" />
        ...
      <joint name="SPI1" .... >
        <joint name="SPI2" .... >
          <joint name="SPI3" .... >
            <joint name="LCLR" .... >
              ...
            </joint>
          </joint>
        </joint>
      </joint>
    </skeleton>
  </avatar>
</avatarStaticData>
```

4. Animgen Configuration Data File Format

The `config.xml` files contain the settings for controlling many of the aspects of the synthetic signing generated by Animgen, and is loaded by Animgen when processing a SiGML file to produce animation. The file defines timings, signing space, constraints, trajectories, hand shapes, constants, repetitions, and rest poses.

For example, the following code defines a handshake for a fist with the index finger extended.

```
<handshapes>
  <finger2
    specialbends="0000"
    ordinarybends="4440"
    extendedfingers="2"
    class="fist"
  />
</handshapes>
```

Each finger bending consists of 4 numbers, representing respectively the bends at the first, second, and third joints, and the splay angle. For each of these, 0 represents the value when the joint is not bent and 4 is its maximum bending. Each handshape has two different finger bendings: "specialbends" is the bending of the extended fingers (e.g. the index finger for the finger2 handshape) and "ordinarybends" for the other fingers. The thumb is not described here. "extendedfingers" is the set of extended fingers (which includes the thumb for some handshapes).

The ARPToolkit provides facilities to interactively set values for hand shapes in the `config.xml` file, reloading the modified file and displaying the changed handshape in real time.

Signing space for the avatar is defined in terms of the avatar's dimensions such as arm lengths and feature points on the torso.

```
<signingspace
  horiz_spacing = "0.8"
  vert_spacing = "0.25"
  inout_spacing = "0.15"
  signspacesitesize = "1.2"
  fan = "0.6" curve = "1"
  nearbelly = "0.10"
  torsositesize = "0.10"
  neckheight = "0.02"
/>
```

Before processing a SiGML file Animgen will first load a `config.xml` file from a directory common to all avatars. It will then load an avatar specific `config.xml` file that may contain alternative settings that will override those in the common file. A typical example of this would be settings for hand shapes, where variations in bone sizes between skeletons may affect hand shapes.

5. Nonmanuals File Format

The purpose of the `nonmanuals.xml` file is to define how each SiGML/HNS nonmanual feature is implemented using the avatar's morph targets. It maps the standard names of nonmanuals used in SiGML/HNS to the names of the morph targets in the main avatar definition file, or to parallel and sequential sets of these morph targets. The mapping also includes durations and trajectories (timings) for these nonmanuals. The file also includes mappings from Sampa.

5.1 Examples.

A mapping from a SiGML name for a mouth gesture to an avatar's morph names:

```
<mouth_gesture sigmlName="D01">
  <parmorph>
    <morph name="eee" amount="0.6" timing="x m t s m l x"/>
    <morph name="ulpr" amount="0.2" timing="x m t s m l x"/>
    <morph name="ulpl" amount="0.2" timing="x m t s m l x"/>
  </parmorph>
</mouth_gesture>
```

For the SiGML mouth gesture D01 the gesture comprises the morphs “eee” with an amount of 0.6, “ulpr” with an amount of 0.2, and “ulpl” with an amount of 0.2. Enclosing all three in the <parmorph> </parmorph> element indicates that these should be combined in parallel. All parallel combinations of morphs must have the same timing, with the optional “x”, in this case, at each end indicating that this gesture should be adjusted to last the same length of time as the manual gesture that it accompanies.

A mapping from Sampa to an avatar's morph names:

```
<sampa phonemes="O_I:">
  <morph name="ooo" timing="m t - m t"/>
  <morph name="eee" timing="m t m m t"/>
</sampa>
```

Here “O_I:” represents a diphthong which is mapped to two morphs, “ooo” and “eee”, performed in sequence, each with a different timing.

Non-facial nonmanuals. These are animations of the head, spine, and shoulders that are expressed as “pseudomorphs” in SiGML, but are processed by Animgen into bone animations.

```
<head_movement sigmlName="NO">
  <morph name="HTLF" amount="0.03" timing="m t - f l"/>
  <morph name="HTLF" amount="-0.03" timing="m t - f l"/>
  <morph name="HTLF" amount="0.03" timing="m t - f l"/>
  <morph name="HTLF" amount="-0.03" timing="m t - f l"/>
  <morph name="HTLF" amount="0.03" timing="m t - f l"/>
  <morph name="HTLF" amount="-0.03" timing="m t - f l"/>
</head_movement>
```

This produces a set of sequential bone movements of the head from left to right - “NO”.

5.2 Durations and Trajectories

The timing attribute for each morph is a sequence of up to 7 tokens, each with codes that map to constants defined in the config.xml file, with the following purpose:

- 1) Whether the morph is anchored to the start of the interval during which it is played.
- 2) The attack time (the time spent ramping up from zero to the full amount).
- 3) The attack trajectory (the manner in which it approaches the full amount).
- 4) The sustain time (the time spent holding the morph at its full amount).
- 5) The release time (the time spent ramping down to zero).
- 6) The release trajectory (the manner in which it ramps down to zero).
- 7) Whether the morph is anchored to the end of the interval during which it is played.

The first and last token is either ‘x’ (anchored) or ‘e’ (elastic). These tokens can be omitted, and default to ‘x’ and ‘e’ respectively. Each time is either a real number of seconds, or one of the following tokens:

```
f fast
m medium speed
s slow
- zero
```

Each trajectory is one of the tokens “t” (targetted) or “l” (lax). The targetted trajectory makes a greater acceleration and deceleration towards its endpoint. Typically one would use “t” for everything except the release trajectory of the last morph.

6. Conclusion

The requirements to enable an avatar to perform deaf signing in the UEA JASigning software are essentially in addition to the standard specification for any avatar that can be animated. The only additions to the standard specification are the extra morph targets specific to deaf signing. The representation of the standard data can be converted to the format used in the avatardef.arp file already described. The other additional data required for signing is held in the asd.xml, config.xml, and nonmanuals.xml files. We believe these additions to the requirements for a standard virtual human character definition will be necessary in any system that synthesises authentic animated sign language.

7. Acknowledgements

We acknowledge with gratitude that the work described here has been partially funded under the European Union’s 7th Framework Programme, through the Dicta-Sign project (grant 231135).

References

- Elliott, R., Glauert, J.R.W., Jennings, V., and Kennaway, J.R., “An Overview of the SiGML Notation and SiGML-Signing Software System”, In Fourth International Conference on Language Resources and Evaluation, LREC 2004, Edited by Streiter, O. and Vettori, C., Lisbon, Portugal, pp. 98-104, 2004.
- Elliott, R., Glauert, J.R.W., Kennaway, J.R., Marshall, I., and Safar, E., “Linguistic modelling and language processing technologies for avatar-based sign language presentation”, Universal Access in the Information Society, vol. 6, no. 4, pp. 375-391, 2007.
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J., et al. “Hamburg Notation System for Sign Languages—An Introductory Guide”, International Studies on Sign Language and the Communication of the Deaf (5). Institute of German Sign Language and Communication of the Deaf, University of Hamburg, Hamburg, 1989.
- Kennaway, J.R., Glauert, J.R.W., and Zwitserlood, I., “Providing Signed Content on the Internet by Synthesized Animation”, ACM Transactions on Computer Human Interaction, vol. 14, 3, no. 15, pp. 1-29, 2007.
- Hanke, T., “HamNoSys representing sign language data in language resources and language processing contexts”, In Fourth International Conference on Language Resources and Evaluation, LREC 2004, Edited by Streiter, O. and Vettori, C., Lisbon, Portugal, pp. 1–6, 2004.
- [ARP] <http://vh.cmp.uea.ac.uk/index.php/ARP>

Adding value to, and extracting of value from, a signed language corpus through secondary processing: implications for annotation schemas and corpus creation

Trevor Johnston

Macquarie University

Sydney, Australia

E-mail: trevor.johnston@mq.edu.au

Abstract

A basic signed language (SL) corpus is created through primary processing of video recordings using multi-media annotation software. Primary processing entails the tokenization and identification of SL units. For the purposes of linguistic research a corpus also needs secondary processing. Secondary processing entails appending tags for specific linguistic features to primary annotations. I draw on the experience from the Auslan corpus project to describe how primary and secondary processing can be used in corpus-based SL research. In particular, I show how the tier structure of ELAN can be used to tag SL units in a variety of ways, and how this information can be used to glean new information from the corpus which can then be added as new annotations to the corpus. Value-adding by principled and systematic primary and secondary processing of digital recordings is thus not only essential for corpus creation ('machine-readability'), it also enables further enriching of the corpus so that even more value can be extracted. I conclude by discussing the implications for annotation software and standardized annotation schemas used in the creation of SL corpora.

The case for SL corpus linguistics

There are many arguments that have long been advanced in support of corpus-based language description and linguistic research and they all apply equally well to SLs. There is no time to repeat them here even if they do go to the very heart of what it is linguists treat as (sufficient) evidence for a claim about the grammar of a language. Suffice it to say, however, that I take them as strong arguments in favor of basing descriptive and theoretical linguistics on how people use a particular language, and not on their intuitions or judgments (at least, not alone). However, there are several additional reasons why corpora are particularly important in SL research, and some of them are unique to this field of linguistics. They do bear repeating, e.g. see Johnston & Schembri (2010).

SLs are languages of minority communities that rarely have any real geographical centre, apart from perhaps residential schools for the deaf or deaf clubs. SLs experience interrupted inter-generational transmission for all but a tiny minority of users and thus have few native users. SLs have no dedicated or widely used written form, nor long history of being used in education.

These facts create two major problems for SL researchers. First, intuitions may be less useful in language description work in SL-using communities, all of which have been characterized as displaying high degrees of variation in both lexis and grammar. Moreover, users sometimes appear to lack sets of shared linguistic norms that are often found in stable language communities, especially those with literacy and standard varieties used in education. This variability means there may be little consensus on phonological or grammatical typicality, markedness or acceptability among users. The practice in SL linguistics of relying on the intuitions of a small number of informants can thus be seen as problematic (even if one was to give high evidential status to intuitions and/or grammaticality judgments in the first in-

stance). Second, the representation of SL utterances using written glosses has meant that primary data have remained essentially inaccessible to other researchers and consequently unavailable for meaningful peer review.

In short, there is a particular need for SL recordings which can be processed into language corpora in order to empirically ground our understanding of the structure, use, acquisition and learning of SLs, and to test claims or hypotheses about their grammars. Without corpora, one risks basing educational interventions and interpreting training, the design of automatic SL processing or recognition systems, and even linguistic theory itself on descriptions of SLs that may be inadequate.

Language processing and corpus linguistics

In the history of SL research almost no extended SL texts of any kind have been created, either by glossing or by using a dedicated notation system, that could in turn be digitized, read by computer and further processed.

With recent advances in digital recording technology, computing, and multimedia annotation software, the way in which recordings of face-to-face language could be best processed to create corpora for the purposes of linguistic analysis has been transformed (cf., Beal, Corrigan & Moisl, 2007). For instance, the source text can now remain the primary data itself, rather than being necessarily replaced by its representation in a transcription to which annotations were subsequently appended. This has made the creation of SL corpora feasible. One of a number of multimedia annotation software programs suitable for use by SL researchers wishing to create corpora is called ELAN (Max Planck Institute for Psycholinguistics Language Archiving Technology Group, 2009).

A minimalist corpus: primary processing

A basic signed language reference corpus is created

through primary processing of the raw video recordings in an archive using multi-media annotation software, e.g. ELAN. Primary processing entails the tokenization and identification of signed units. This can be achieved by ensuring that conventional linguistic units and types are systematically and consistently identified with invariant and unique sign identifiers (or “IDglosses”). Consistency in type/token identification is the key requirement for ensuring that a SL corpus is machine-readable for the purposes of linguistic research (see Johnston, 2010).

This is achieved by corpus annotators adhering to set protocols and schemas with respect to the classification and identification of sign types and the assignment of IDglosses to fully lexical signs. The Auslan corpus project has developed such a set of guidelines and other SL corpus projects are in the process of developing their own.¹ In SL corpora, attention must be paid to distinguishing between fully-lexical signs and partly-lexical signs (both content signs and grammatical signs) and gestures (both manual and non-manual).

A minimalist corpus also usually involves the addition of a time aligned parallel translation into the working majority spoken language. Indeed, in some very basic corpora the only annotation may be a parallel translation, grossly time-aligned to the source media.

Just on the basis of primary processing of a corpus, it is possible to glean valuable information on sign tokens, sign types, or signs by IDgloss, e.g. number, frequency, duration, and concordance/collocation patterns. It is even possible to conduct preliminary and tentative grammatical analyses, by locating segments of the primary text that co-occur with particular constructions in the translated parallel text.

Before turning to secondary processes, I will briefly exemplify how these primary annotations can be used to extract this type of information in the ELAN search routines. However, partly because of space constraints in this paper and time constraints in the presentation, I will only be able to discuss frequency and collocation.

IDgloss frequency

Selecting within the ELAN menu thus: > Search > Single Layer Search, one defines the search domain (keeping separate left hand dominant from right hand dominant signers), selects the mode (annotation, regular expression), selects the tier (IDgloss) and specifies the search. In this case, .+ for “any text”. There are 41,842 hits in the result of which approximately 10% are represented in the top 10 most frequent IDglosses (Figure 1).

Substring match searches can be used to specify the beginning of an annotation string (such as ^PT “begins with PT”, ^DS “begins with DS” and so on). In this way, one can exploit the glossing conventions for partly-lexical and non-lexical signs and gestures to search for these types of signs by general type (e.g., ^PT or “a pointing sign”) or more specifically (e.g.

^PT:PRO1SG(7) or “first person singular pointing sign made with and index finger and extended thumb”).

Annotation	Percentage	Count
PT	3.63%	1517
PT:PRO1sg	2.65%	1107
G:well	1.81%	756
DEAF	1.47%	615
LOOK	1.41%	589
BOY	1.19%	499
SAME	1.11%	464
HAVE	1.03%	430
PT:PRO3sg	0.98%	408
THINK	0.77%	324

Figure 1: Frequency view of IDgloss search results²

Using this method, it was established by searching the first ‘minimalist’ annotated texts in the Auslan corpus dating from 2006-07 that approximately 11% of all signs in the corpus were points, 7% were gestures, and 10% were depicting signs (i.e. up to almost 30% of all signs produced were either non-lexical or partly-lexical signs). Interestingly, as the corpus has grown, from 10,000 to 60,000 sign tokens, these relative proportions have changed little.³

IDgloss (fully-lexical signs only) frequency

Using the same procedure as in the previous search but with the search text specified as:

^[^QPT\E][^QDS\E][^QFS\E][^QG:\E]

for “begins with any text except PT (point), DS (depicting sign), FS (fingerspelling), or G (gesture)” (in other words, “find all lexical IDglosses.”) yields all lexical signs. There are 25,750 hits in the result, but only the top 10 are displayed in Figure 2.

Annotation	Percentage	Count
DEAF	2.39%	615
LOOK	2.29%	589
BOY	1.94%	499
SAME	1.80%	464
HAVE	1.67%	430
THINK	1.26%	324
NOTHING	1.24%	320
GOOD	1.22%	315
WHAT	1.11%	287
WHY	1.08%	279

Figure 2: Frequency view of lexical sign search hits

Collocation and frequency

Using the same procedure as in the previous search but with search type specified as *n-gram over annotations* and the search text as # think (for “any two sequential annotations, the second of which is THINK”), the results (out of 330 hits) are displayed in Figure 3. (Once again, the table only displays the top 10 hits.)

² Signs glossed simply as PT have yet to be further specified.

³ The aim is to expand the corpus to 100,000 sign tokens by the end of 2010 and to double that number again by 2012 by increasing the number of annotated digital movies from the current 201 clip to around 500. There are more than 1,200 movie files in the corpus.

¹ The Auslan annotation guidelines can be downloaded from <http://www.auslan.org.au/about/annotations/>

Annotation	Percentage	Count
PT THINK	19.50%	63
PT:PRO1sg THINK	15.48%	50
NOT THINK	2.48%	8
NEVER THINK	2.17%	7
WHAT THINK	1.86%	6
THINK	1.86%	6
LOOK THINK	1.55%	5
? THINK	1.55%	5
PT:PRO3sg THINK	1.24%	4
PEOPLE THINK	1.24%	4

Figure 3: Frequency view of signs preceding THINK⁴

These searches are only possible because of distinctions made in the IDglossing between type and token, and between sign sub-types. However, the real efficacy of this type of annotation schema becomes best seen if we look at its place in secondary processing.

A value-added corpus: secondary processing

For the purposes of conducting detailed linguistic research a corpus also needs to undergo secondary processing.

Secondary processing entails appending information to annotations created in primary processing. These secondary annotations (or ‘tags’) add specific phonological, morphological, syntactic, semantic, pragmatic or discourse information about linguistic forms, depending on the purpose of the analysis. In ELAN the tags are distributed over multiple tiers, each dedicated to a certain type of tag. Once again protocols and schemas need to be implemented to ensure that the tags used are drawn from a limited or controlled vocabulary of values and that they are applied to the primary annotations in a consistent manner. These too are covered in the annotation guidelines for the Auslan corpus.

Secondary processing enables one to extract far more sophisticated frequency statistics for any annotation (IDgloss or linguistic tag) and to specify and identify the environments in which they occur in greater detail. For example, ELAN searches can be constrained by specifying aligned or overlapping values on as many as two other tiers for any specified annotation or string of up to three annotation values. In addition, multiple annotation files can be specified as the search domain. These can be selected manually or automatically based on metadata values such as age, gender, region, text type, etc.

The analysis of the search results can be partially done though examining ELAN’s search results directly or by exporting them in various formats. For example, once the matches have been computed they are displayed in either concordance or in frequency views in the ELAN search dialogue box. Both of these data types can then be exported for further processing in various databases or

corpus analysis software programs.

With respect to the Auslan corpus, a number of studies are now underway using texts that have been enriched with secondary annotations, be they formational (palm orientation, handshape, sign location and/or sign directionality), lexico-grammatical (grammatical class, argument structure, semantic roles, ‘PRO-drop’), and ‘utterance’ level (clause boundaries, constructed action).

I now describe the procedure that makes it possible to use secondary annotations in the ELAN search routines to extract interesting and relevant linguistic observations. Once again, due to space and time constraints, I give only a few examples—palm orientation, grammatical class, and clause argument structure—as well as briefly discussing constructed action. I only give example data drawn from subsets of the Auslan corpus. A formal report using corpus-wide and definitive data is not my purpose here.

Palm orientation and pointing signs

Selecting from the ELAN menu Search > Multiple Layer Search, one then defines the search domain, selects the mode and the search tiers (1 IDgloss, 2 orientation), and then specifies the search text: ^PT (“begins with PT”) for the IDgloss and .+ or “any text” for the orientations (d = down, l = left, u = up, r = right, o = other), as well as specifying that both annotations *overlap*. The results in an example subset of 19 eafs have 244 hits (only top 10 displayed, see Figure 4).

Annotation	Percentage	Count
#1 PT:LOC #2 ldl #3	20.08%	49
#1 PT:LOC/PRO3SG #2 ldl #3	14.34%	35
#1 PT:PRO3SG #2 ldl #3	9.02%	22
#1 PT:LOC/PRO3SG #2 #3	7.79%	19
#1 PT:DET #2 ldl #3	7.79%	19
#1 PT:PRO3SG #2 #3	5.74%	14
#1 PT:DET #2 #3	3.69%	9
#1 PT:FBUOY #2 ldl #3	3.28%	8
#1 PT:LOC #2 #3	2.46%	6
#1 PT:PRO3PL #2 ldl #3	1.64%	4

Figure 4: Frequency view of PTs & orientation

Naturally, because of the systematic nature of IDglossing, sign types, be they non-lexical or partly-lexical signs, are able to be filtered through substring search matches to extract more specific hits. For example ^PT:PRO3|PT:PRO2 will find all third *or* second person pronouns (see Table 1).

	PT:LOC	PT:PRO3/PRO2
down	62%	58%
left	38%	38%
other	0%	4%
Total	100%	100%

Table 1: Results specifying for point type

There has been some discussion in the literature about the association of a downward palm orientation in pointing signs that are strongly associated with a location

⁴ The six instances in which no sign precedes THINK are instances in which there has been a switch in hand dominance to the subordinate hand.

(‘here/there’) and/or could be described as demonstratives (‘this/that’), rather than being used simply pronominally. Even though the categorization of points in the Auslan corpus does not correspond neatly to the classes of pronouns, locatives, and demonstratives in traditional grammars, the data to date extracted from the Auslan corpus, of which the data in Table 1 is just an example, does not appear to show an association of a point with a palm turned downwards with at least locative meanings. It remains to be seen what a large reference corpus will show.

Lexical frequency by grammatical class

In the Auslan corpus there are annotations that assign grammatical class membership to sign tokens in context. In the multi-file multi-tier search dialogue lexical IDglosses can thus be constrained as co-occurring (overlapping) with grammatical class tags. The results can be view in frequency view and/or exported to databases for further sorting. Example results in Table 2 are based on two specific IDglosses, as shown:

	FINISH-FIVE %	FINISH-GOOD %
Adjective	5.10	0
Adverb	5.10	17.14
Auxiliary	36.74	31.43
Conjunction	2.04	5.71
Discourse marker	6.12	8.57
Interactive	1.02	0
Noun	3.06	2.86
Predicate	6.12	14.29
Unsure	1.02	0
Verb	33.68	20.0
Total	100	100

Table 2: The lexical frequency of two ‘verbs’ in the semantic area ‘finish’ specified by grammatical class.

The only major large lexical frequency study of any SL (McKee & Kennedy, 2006) did not, strictly speaking, take grammatical class formally into consideration in so far as it was assumed that the grammatical class of the English glosses used for each sign token accurately reflected each token’s use *in situ*. In reality, glosses usually name the most frequent use of a sign, not its actual use in context.

Clause argument structure

In the Auslan corpus there are annotations that delimit clause boundaries. IDglosses are tagged for their status as arguments of the verb which is also tagged (e.g. as process, utterance or enactment). After merging tier annotations which combines these clause tags, it is relatively easily to identify and quantify clause construction types. For example, from the ELAN menu, > Tier > Merge Tiers, one selects tiers to merge (select ‘concatenate’). View annotation statistics and select the newly created merged tier. Export to databases for further processing if necessary (a sample result from one file is

shown in Table 3).

In the Auslan corpus there are also annotations that tag the identified overt arguments for their semantic role in the clause (e.g. as agent, patient, experiencer, etc.). By first merging the argument tag tier with the semantic role tier, before merging the result with the clause annotation tier, it is possible to extract richer data (Table 4).

Clause construction by order of overt arguments	#
V	27
A V	7
A1 V A2	6
V A	6
A1 A2	4
A	3

Table 3: Frequency of clause construction types

Clause construction by order of overt arguments	#
V (PROCESS)	27
A (AGENT) V (PROCESS)	6
A1 (AGENT) V (PROCESS) A2 (PATIENT)	4
A1 (CARRIER) A2 (ATTRIBUTE)	3
V (PROCESS) A (PATIENT)	3
A (ATTRIBUTE)	2
A (EXPERIENCER) V (PROCESS)	1
A (UTTERANCE)	1
A1 (AGENT) V (PROCESS) A2 (GOAL)	1
A1 (ENTITY) A2 (LOCATION)	1
A1 (EXPERIENCER) V (PROCESS) A2 (SOURCE)	1
V (PROCESS) A (ENTITY)	1
V (PROCESS) A (LOCATION)	1
V (PROCESS) A (UTTERANCE)	1

Table 4: Frequency of clause construction types specified for semantic role of argument

The data in Table 4 are only indicative of the type of information that can be extracted regarding clause structure based on secondary processing and are only taken from a single annotation file. Of 201 movie clips that have currently undergone primary and secondary processing, less than 10 have also been annotated for clause boundaries, overt arguments *and* semantic roles.

Though the range of clause construction types and the possible alignments of semantic roles to various argument positions commonly found in Auslan already appears much wider than that shown in the example file above, it is far too early to draw firm conclusions. A formal report of this data and its possible significance in describing grammatical structure in Auslan is not planned until the clause annotation set reaches at least 50 files and/or several thousand clauses.

Verb type by modification and by CA co-occurrence

In the Auslan corpus there are annotations that delimit periods of constructed action (CA). Multi-file searches constrained by values over three tiers can thus be based

on the co-occurrence of tags for grammatical class (verb type), spatial modification (present or absent), and constructed action (present or absent). The results can then be exported to database programs. Relevant metadata regarding text type, age, and region, for example, can be easily appended to each token/hit in the exported data as ELAN automatically appends the file name source of each. This can then be run through statistical programs to test for factor interaction and significance.

Further value-adding: tertiary processing

The observations relating to single sign or multi-sign constructions extracted from a corpus using the procedures exemplified above are valuable in their own right. However, there is another, perhaps overlooked benefit to this type of SL corpus linguistics. The findings extracted from a corpus can themselves, in turn, be fed back into the corpus annotations, as part of an augmented secondary processing. They can then be used to generate yet further observations. I refer to this augmenting process here as *tertiary processing*.

For example, the very identification of a set of extremely high frequency lexical verbs in Auslan was only made possible because the corpus was not only annotated, but annotated in a systematic way that identified lemmas and, later, their grammatical class in context. The lexical frequency of sign types was then able to be added to IDglosses, filtered by grammatical class, as a frequency tag. In other words, researchers were able to find all instances of an IDgloss with a given grammatical class tag and replace that gloss with a tag signifying the lexical frequency of that sign (e.g., VHF for ‘very high frequency’, HF for ‘high frequency’, and LF for ‘low frequency’).

Augmentation of an exported data in this way can be done semi-automatically in database programs by filtering records and adding tags in fields for the relevant subset of records. The tag can then be added as another factor in subsequent re-evaluation of the data.

Inserting these tags into the ELAN file itself is worthwhile because there is currently a three-tier limit for simultaneous constraints in multi-tier multi-file searches. This means that any constraint which is itself the product of condition matching over two or three tiers, cannot itself be constrained further. By inserting such a derived value into the ELAN annotation file, this automatically means that this value can be used freeing the other query tiers to specify additional constraints.

Though the replacement or tagging process is not automatic within ELAN, there are workarounds. They take some time to do but since they need to be done only once and the results are always available for use, they are worth the effort (but see implications below). For example, the IDgloss tier can be copied or filtered to a new tier designed to hold the frequency tags. Then, the glosses on the derived tier are searched and replaced with the appropriate tag according to the lexical frequency by grammatical class table that has been generated by prior analysis. This can be done across multiple

files, if not the entire corpus, in one operation. The workflow moves from the very high frequency signs to low frequency signs, as the very high or high frequency sign types are relatively few in number. (Of course, there are many tokens of these types!)

In other words, first with respect to high frequency signs, all IDglosses for a particular lemma are replaced with the same tag on the assumption they are all of the same grammatical class as the most frequent member. Then the remaining members of different grammatical classes—a much smaller set—are identified and the tag changed accordingly. With respect to low frequency signs, they can all be tagged as low frequency in one single universal search and replace: “find all annotations on the relevant tier which are *not* VHF or HF and replace with LF.”

Similarly, as mentioned above, it is relatively easy to extract occurrences of signs that co-occur with periods of constructed action in a text. Tags for co-occurrence can then be added to the IDglosses (according to grammatical class).

Both frequency and CA co-occurrence information have been incorporated into a subset of the Auslan corpus in ways described above and were used in the recent study by de Beuzeville, Johnston and Schembri (2009) on the spatial modification of verbs in the Auslan corpus. This study examined the frequency and linguistic environments of verb modification with a view to assessing if spatial modification to signal these roles was obligatory in the language. The spatial modification of verb signs in SLs has traditionally been explained as a grammatical system marking subject and object roles (e.g., Sandler & Lillo-Martin, 2006), similar to obligatory subject marking in English (e.g., third person singular *-s* in *he walks*).

The Auslan study found that the modifications were not obligatory, were strongly associated with a very small number of high frequency verbs, and tended to co-occur in specific linguistic environments (e.g., co-occurrence with constructed action). The authors suggested that these observations would not be expected under the traditional grammatical account of spatial modification and are more in keeping with an analysis that sees the phenomenon reflecting, in part, the fusion of gestural pointing into the articulation of lexical verbs, as suggested by Liddell (2003).

It is anticipated that similar procedures as those described here will integrate derived clause argument structure patterns into tags added to clause annotations within the Auslan corpus. The patterning of clause chains (e.g. with overt or elided arguments, or with certain verb/argument sequences) and their interaction with verb modification, depicting signs, constructed action (as well as other linguistic variables) may then become identifiable and amenable to quantification and further analysis.

Standardizing annotation schemas

The type of investigations of the Auslan corpus that we have briefly illustrated here have only been made possi-

ble because of the distinctions made in the IDglossing between types and tokens, and between sub-types of signs. The way these distinctions are coded in the in the IDglosses are described in detail elsewhere (e.g. in corpus annotation guidelines⁵). The primary, secondary and even tertiary processing of language corpora is extremely time consuming work. However, the results more than justify the effort expended in adding value to raw language recordings — recordings which would otherwise be of limited use — in this way.

The international standardization of annotation practice, protocols or schemas is highly desirable. Indeed, at the level of primary processing this should be a high priority. At the level of secondary processing, however, there is much more room for flexibility as the aims of various research teams can be very different, each perhaps requiring its own dedicated secondary tags. Standardization, in so far as it is possible, will certainly enable the corpus-based comparative analysis of SLs to be undertaken.

Within a give SL corpus, however, there is really no option: standardization in terms of systematicity and consistency is mandatory. Only in this way can annotations create machine-readable SL texts that can be searched rapidly and with great precision. The results can then be further processed for statistical significance and interaction, or, just as importantly, the hits further examined individually *in the media context* to assist in the determination of their semiotic or linguistic significance.

Implications for annotation software

From the discussion above, it will be evident that the steps needed to conduct some searches or data exports are in need of automatization. For instance, preparations for some multi-tier pattern match searches, on the one hand, or merging information coded on separate tiers, on the other, are ad hoc and time consuming. External plug-in scripts are one solution. However, fully integrated improved program functionality is preferable as it means all researchers using the same software have the same functionality available.

With respect to ELAN, for example, these scripts or routines would enable one to automatically create, copy or merge certain tiers in multiple annotation files of the same type; automatically look up an alternative value for an annotation in a table and substitute that value for the annotation on a particular tier in multiple annotation files; or automatically place a specified value in an empty annotation field which is the result of a hit specifying the overlap of two annotations of two other tiers (independent or otherwise).

Search functionality also needs to be improved so that more than three tiers may be specified in constrained pattern matching. Most importantly, the co-occurrence (or non-occurrence) of two given annotations within the

time delimitation of a single annotation on another tier should able to be specified as a search condition.

Conclusion

The creation of SL corpora as corpora in the modern sense involves more than recording, digitizing, editing, cataloguing and archiving video texts. Corpus creation must also involve the transformation of archived material into something which is machine-readable by the principled application of annotation procedures that make optimal use of new digital technologies. By adding value to a corpus through systematic and principled primary and secondary processing, it is possible to extract the true value inherent in a linguistic corpus.

Acknowledgements

Research towards this paper was funded by an Australian Research Council Discovery Project grant (#DP1094572) awarded to Trevor Johnston.

References

- Beal, J. C., Corrigan, K. P. & Moisl, H. L. 2007. "Taming digital voices and texts: Models and methods for handling unconventional synchronic corpora". In J. C. Beal, K. P. Corrigan & H. L. Moisl (Eds), *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases*. New York: Palgrave Macmillan, 1-16.
- De Beuzeville, L., Johnston, T. & Schembri, A. 2009. "The use of space with lexical verbs in Auslan: A corpus-based investigation". *Sign Language & Linguistics*, 12 (1), 53-82.
- Johnston, T. 2010. From archive to corpus: transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1), 104-129..
- Johnston, T., & Schembri, A. 2010. Variation, lexicalization and grammaticalization in signed languages. *Langage et Société*, 131(mars 2010), 19-35.
- Liddell, S. K. 2003. *Grammar, Gesture, and Meaning in American Sign Language*. Cambridge: Cambridge University Press.
- McKee, D., & Kennedy, G. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4), 372-390.
- MPI/LAT Technical Group: (Head) Wittenburg, P., (Team members) Auer, E., Broeder, D., Gardellini, M., Kemps-Snijders, M. et al. 2009. *EUDICO Linguistic Annotator (ELAN)* (Version 3.8). Nijmegen, Netherlands: Max Plank Institute for Psycholinguistics: Technical Group (Language Archiving Technology). Available at: <http://www.lat-mpi.eu/tools/elan/>.
- Sandler, W., & Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge: Cambridge University Press.

⁵ The Auslan annotation guidelines can be downloaded from <http://www.auslan.org.au/about/annotations/>

New features in synthesis of sign language addressing non-manual component

Zdeněk Krňoul

University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic
zdkrnoul@kky.zcu.cz

Abstract

A sign language synthesis system converts previously noted signs into the computer animation. The animation is created using a specially designed 3D model of the human figure and algorithms transferring the sign to movements of the model. In principle the sign language contains both the non-manual component (shape and movement of hands) and the non-manual component (facial movements, etc.). Notation of the non-manual component was not yet sufficiently explored in terms of an automatic conversion to the animation. In the article we describe both notation methodology of the non-manual component and technical aspects for conversion of symbols to movements of the animation model. In addition an appropriate animation method for the 3D shape of face is assumed. The result is an extended notation supplementing notation of the manual component with the non-manual component. The extended notation preserves the feasibility of an automatic conversion and keeps the original level of generality. In connection with the methodology we present the notations of the basic types of non-manual components of the Czech sign language.

1. Introduction

The sign language synthesis system including 3D human figure, complete animation of arms as well as movements of body, head, and facial gestures is a promising usage of computer technology to reduce communication difficulties for deaf people. The sign language synthesis system is a part of complex systems translating the text to sign language as virtual interpreters, signing tutors, sign language dictionaries and others. For linguistic research, a symbol based synthesis system provides the immediate feedback, verification of entered notations, etc. The research on non-manual signals (NMS), the non-manual component of Czech Sign Language (CSL), uses the SignWriting notation system (SW). However for the manual component, the Czech sign speech synthesis system uses the Hamburg Sign Language Notation system (HNS). The notation method in order to transform NMS to 3D animation has not been defined yet.

NMS are parts of the sign language as the speech of a spoken language is not just expressed words and grammar. There are signs distinguishable only by the NMS and *the specific signs* without the manual component. NMS has at least six different roles (Bridges and Metzger, 1996). Symbolic notation can be used primarily for lexical, grammatical markers, conversation regulators, non-manual modifiers. For the mouth pictures, we can directly use letters of the alphabet instead of symbols (Elliott et al., 2004). Movements of other parts of the face, head and chest should be noted individually. There are already sign language synthesis systems reanimating a data record of speaker of the sign language or systems controlled by a symbolic entry (Elliott et al., 2004; Krňoul et al., 2008). Initial interest was directed to the accurate and realistic animation of shapes and movements of the hands. An extension of the synthesis system involves new algorithms for conversion of NMS to 3D animation. The methodology provides universal notation of the non-manual components of sign languages and guarantees automatic processing of it by a computer system.

Section 2 introduces the concept of notation of NMS by

HNS and includes the notation of the basic types of non-manual components of the Czech sign language. Technical aspects of conversion to 3D animation are discussed in Section 3 and Section 4 is the conclusions.

2. Notation of Non-manual Signals

Well-known sign language notation systems are Stokoe, SignWriting (SW), HamNoSys (HNS) (Stokoe et al., 1976; Rosenberg, 1995; Schmaling and Hanke, 2001). In terms of non-manual signals (NMS) SW seems to be the most complex notation system. Notation of NMS has to include not only constructions for facial expressions but also movements of upper parts of the body, head and eyes. Minimal observable actions in the face are also in the detail treated by action units (AU) of Facial Action Coding System (FASC) (Ekman et al., 2002). HNS has very a detailed notation of the manual component but the non-manual component is only adumbrate. In contrast, the structure of signs in HNS is suitable for computer processing. We have a synthesis system creating 3D animation of the manual component from HNS (Krňoul et al., 2008) and consider the collection of HNS symbols of the version 4.0 to be sufficient enough for this notation purpose.

The position of non-manual component in structure of HNS is depicted in Figure 1. HNS does not have symbols for complex gestures but the gestures can be notated by a couple of symbols. We consider NMS to be expressed by one or more *non-manual actions*. One non-manual action describes *the rotation and movement of joints*, or *the movement in the face*. A general notation form has in following order: *a base symbol* and *control symbols*. Furthermore the base and control symbols can optionally be supplemented by additional auxiliary symbols (modifiers).

2.1. Transformation of Joints

Non-manual action for transformation of the joints can be used for movements of stomach, chest, shoulders, head and eyes (eye gaze). We consider these base symbols: \square \square ∞ . The base symbol \square (shoulder) may be accompanied by the

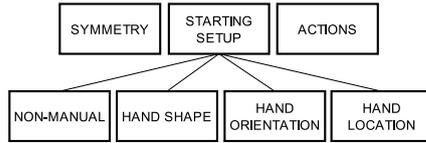


Figure 1: The position of the non-manual component defined by HamNoSys.

symbol \cdot (side modifier) restricting the non-manual action to the left or right side of the body.

The control symbols define both rotation and movement of body part represented by the base symbol. For the rotation, we propose to use symbols of *the finger base direction* originally defined for the manual component. There are 18 symbols defined: $\wedge \vee \succ \dots$. All combinations of these symbols and the base symbol cannot be interpreted because the joint limits allow only possible poses of human body. We are trying to preserve the original meaning of these symbols. For the base symbols $\square \equiv$ (stomach and chest), the rotation establishes turning in the direction from the body. Hands and head are turning with the chest. In this case, the control symbol \triangle determines the neutral pose, \triangleright is turning on right etc.

The meaning of rotation is not such obvious for the symbol \square (shoulders). A rotation of shoulders is evaluated for the dominant hand. We consider four main directions: $\wedge \vee \triangle \times$ (up/down, forward/backward). The shoulder of the non-dominant hand is determined the left/right symmetry. For the head, the control symbols determine the direction of the nose and eyes are rotated with head (the direction of view). For example, one base and one control symbol is used for head turning to the right: $\square \triangle$. We can optionally add one more control symbol to notate rotation of joints more precisely. The joints are turned in the same meaning as *the palm orientation*. For example, $\square \triangle \circ$; it describes nose direction forward but chin is rotated on right. This optional specifies the rotation especially for chest and head where we expect still a tilt of the joint.

The base symbol may be in combination with symbols for movements as well. The movement will be carried out in the base pose, or in the noted direction. For example, NMS for a head moving from side to side is $\square \triangle \uparrow$ or only $\square \uparrow$. Rotation of jaw and lower teeth is not considered as movement of the joint but rather as part of a movement in the face. In addition, we propose eye contact with the hands as well. Basic notation is the following short combination of the base symbol and one modifier: ∞^x .

2.2. Movements in Face

Movements in the face are changes in the shape or the position of a forehead, eyebrows, the area around the eyes, eyelids, nose, cheeks, chin (skin around the chin), and mouth. The base symbols $\infty \wedge \vee \triangleright \triangleleft \smile \frown$ identify parts of the face (locations) that will be changed. The base symbols $\infty \sim \infty \wedge \vee \triangleright \triangleleft \smile$ can be optionally noted in combination with the side modifiers. We consider the following modifiers, $\cdot \sim \sim$, which reduce the base symbol to more detailed parts of the face.

The meaning of these modifiers is the same as for locations of the manual component. We can specify the non-manual action for left, or right half of mouth, cheeks, eyebrows and eyelids, and the upper or lower lip, the eyelid, or teeth. If these modifies are not used then non-manual action will be performed for both the left and right half of the face, or both the upper and lower lip, or the eye lid.

For movements in the face, noted control symbol determines elementary movement whereby the shape of the face is deformed. For this purpose, we propose to use symbols for straight movements: $\uparrow \nearrow \rightarrow \dots$. These symbols determine 18 elementary movements to control the shift of non-manual actions in 3D space. Furthermore the control symbol may be supplemented with the following modifiers: $\circ \cdot \ast \neg$. The size of the non-manual action can be distinguished in three levels: normal, small and large. A modality of the movement may be normal, fast, or slow. For example, "eyebrows go down and near" $\sim \wedge$, or "little inflation of the right cheek" $\triangleright \cdot \rightarrow$.

2.3. Additional Movements of Mouth

For the mouth, we have three base symbols: $\ominus \omin� \omin�$. These symbols identify the part of the mouth which will be moved. Furthermore, modifiers allow us to specify more detailed positions. Basically the control symbol may have the same use as the symbols for *the straight movements* for the other parts of the face. We assume meaning of these symbols as direction of a contraction of facial muscles around the mouth as well as a complex articulatory movement.

For the shape of lips, we use the base symbol $\omin�$. The non-manual action describes both mouthing (mouth pictures) and mouth gestures. The shapes of lips for mouthing have already been investigated. The studies confirm the use a combination of three or four key shapes: lip opening, lip protrusion, lip raising, and stretching the lips to the side. A combination of these key shapes allows us to note any form of mouthing. For this purpose, we propose simply to use four symbols for straight movements: lip opening \uparrow , lip raising \nearrow , lip protrusion \triangleleft and lip stretch \rightarrow . The remaining symbols from this group determine other elementary actions applicable for the upper and lower lip or the left and right side of the mouth. There are the directions: up \uparrow , down \downarrow , diagonally up \nearrow , diagonally down \searrow , etc. If we do not use the side modifiers then control symbol will have meaning for the right half of the mouth (dominant hand) and the left side will be performed accordance to the left/right symmetry.

For notation of tongue body, we use the base symbol $\omin�$. Again, we can determine its movement in three basic directions: up/down, forward/backward, and left/right. The last base symbol is $\omin�$ (teeth). This symbol is recommended to use only for mouth patterns incorporating "uncovered teeth". The rotation of the chin or lower teeth is implicitly included in the non-manual actions describing the lips or tongue positions and does not need to be noted.

The symbols for the straight movements are not sufficient for precise notation of all mouth patterns. Shapes of the mouth often involve contact lips, teeth and tongue with each other. We assume to use two base symbols and one connec-

tion symbol $\overset{x}{\sim}$ (contact). The notation of such non-manual action is intuitive, for example, the upper lip touches the lower teeth $\overset{x}{\sim}$. A mutual contact of the same symbol has short notation $\overset{x}{\sim}$ rather than $\overset{x}{\sim}$. For slightly opened teeth or lips, we use the shortcut notation of the base and connection symbol $\overset{x}{\sim}$ and the connection symbol $\overset{\circ}{\sim}$ for squeezed lips.

2.4. Usage of Non-manual Actions

The non-manual actions are placed before the manual component, Figure 1. One non-manual action does not need to be explicitly separated from other symbols. However the notation of two or more non-manual actions has to be always enclosed in parentheses. A composition of several non-manual actions allows us to notate more complex non-manual signals (NMS). We assume the same expression as in the manual component. We propose to use two types of the composition. The first type is used for consecutive non-manual actions in time. For this purpose, we have symbols $()$ (parentheses). The non-manual actions are performed consecutively in the order of their notation. The second type is a composition of non-manual actions expressed simultaneously in time. We consider to use the symbols $[]$ (brackets). In this case, all non-manual actions inside produce one fused NMS. Order of non-manual actions is not important.

The combination of these types of composition allows us to note the general NMS. NMS are expressed in parallel to the manual component. Non-manual action begins at the same moment as the movements of the manual component. For example, contact of the lower lip and upper teeth /f/ followed lip protrusion /o/ and simultaneously the head moved slantingly downward and the hand moved in front of the body is noted as: $[(\overset{x}{\sim} \overset{\circ}{\sim} \overset{\circ}{\sim}) \overset{\circ}{\sim} \overset{\circ}{\sim}] \overset{\circ}{\sim} \overset{\circ}{\sim}$. Notations of the basic types of the non-manual components of the Czech sign language originally expressed in SW are summarized in Table 1.

Table 1: Notations of the non-manual component of the Czech sign language, the left column is SignWriting, right column proposed HamNoSys equivalents.

3. Technical Aspects of Non-manual Actions

Technical aspects take into an account problems of the conversion of NMS to the computer animation. Non-manual actions for body joints are expressed by a skeleton structure of the animation model. The principle is same as for the manual component. The animation technique for the face is different. A shape of the face can be created by morph targets, pseudo-muscle actions, control points on the face, or a muscle model (Parke and Waters, 2008). We consider here the shape of the face and tongue as morph targets and the lower teeth as rotation of the rigid body.

3.1. Rules and Rule Actions

The principle of the conversion technique was introduced for the manual component (Křmoul et al., 2008). The schema of the conversion is in Figure 2. This technique automatically carries out the syntactic analysis and creates the parse tree only for the structurally correct entry. Terminal nodes of tree load attributes of particular HNS symbols (descriptors of the symbol). The conversion technique processes the parse tree and reduces its size. Parsing rules join leaf nodes to the parent nodes. Rule actions of the parent nodes integrate attributes from all symbols of the relevant subtree. Next rule actions convert attributes to key frames. The key frames are transformed to the animation frames in accordance with the types and timing of the notated movements.

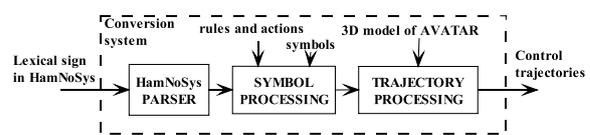


Figure 2: A schema of the conversion system.

The extension of the conversion technique about the new rules and rule actions allows us to accept the input HNS string with non-manual actions. New rules provide the split of parse tree into the manual and non-manual sub-tree, Figure 3. Furthermore rule actions distinguish whether a symbol in the non-manual sub-tree is treated as the base symbol, control symbol, or modifier. The processing of the non-manual sub-tree again takes place in two stages and following order: processing of attributes and processing of animation frames, see Figure 3.

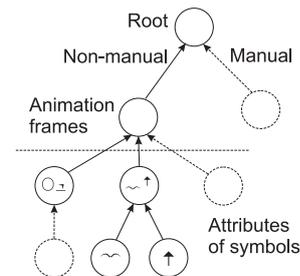


Figure 3: Structure of the parse tree.

The processing of whole non-manual sub-tree precedes the conversion of the manual sub-tree because the animation frames of non-manual actions should be used for the location of the hands. The animation frame consists of two new vectors: *a joint vector* and *a morph vector*. The first one has size $3 \times N$, where N is the number of joints of the skeleton allocated for the non-manual actions. The second one has a size $1 \times M$ and stores the morph weights. M is the number of all considered morph targets.

3.2. Rotations of Joints

The transformation of the non-manual action for rotations of joints is completely solved by the rule actions. The rotation of chest, shoulders, the head and eyes must be eval-

Adapting an Efficient Entry System for Sign Languages

Carlos R. Machado

Serviço Federal de Processamento de Dados - (SERPRO), Brazil
 Associação Software Livre - (ASL), Brazil
 Universidade Federal do Rio Grande do Sul - (UFRGS), Brazil
 machado@softwarelivre.org

Abstract

Building sign language written corpora may, combined with video corpora, provide richer sign language research frameworks. Tools that allow direct sign language writing could increase sign language corpora availability significantly. Here, adaptation of a free efficient computer entry system to allow sign writing is presented.

1. Introduction

Most sign language corpora projects are based in video recordings and their annotation. Although such corpora resources have proven their importance they have almost only gloss notation for annotating the videos and rarely provide direct entry for form aspects of sign languages. Building sign language written corpora may, combined with video corpora linguistics, provide richer sign language research frameworks.

Video corpora annotation resources are not always available for production¹ and post-production² of such videos and later annotation. Providing tools that allow building corpora from direct writing could increase qualified sign language data sets size significantly. Such tools would also improve the meaningfulness of sign language corpora once those would be written by sign language users, mostly deaf. Here an adaptation of an efficient computer entry system to allow sign language direct writing is presented. The following sections introduce predictive writing systems, the tool chosen to be used as a sign language writing tool, the sign language notation and technologies used for the adaptation. Then we report on the current status of the project as well as on future work.

2. Efficient and predictive entry systems

The Human-Computer Interaction research field has achieved important results in a wide variety of input, output and presentation technologies for writing; entry; script, video and audio recognition and many other alternate forms of interaction. There are several entry systems made targeting accessibility and higher efficiency (in general or for specific purposes). Some adopt different layout approaches, others implement inference predictions to speed up writing, novel devices bring approaches that (solely or combined) apply touch, multi-touch, gestures, pressure sensors, video-capture and other techniques. Since inference has become a feature used on a daily basis through mobile phones and mobile computing devices, we present here a discussion on how to implement a novel interface with inference for Sign Language Writing.

¹camera, experienced signers and time

²annotation software, skillful annotation individuals, disk space for video storage and -more- time

2.1. Dasher

Dasher (Ward et al., 2000) is an information-efficient text-entry interface, driven by natural continuous pointing gestures. It is designed to be an alternative entry system when a keyboard is not available or cannot be used. Particularly

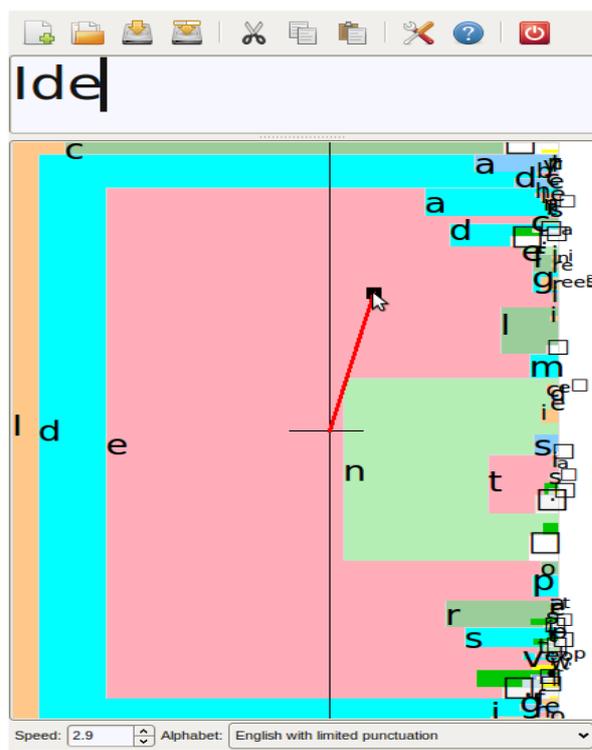


Figure 1: A dasher example. Writing “idea” in English

in cases when a small, mobile device is used to write information or when motionally-impaired computer users may not be able to use a full-sized keyboard.

Amongst the reasons to choose dasher, three are remarkable:

1. Dasher uses continuous gestures movements of a pointing device to “dive” into the symbols.
2. Dasher uses inference based on language model built up from a training text. Using the system increases its

quality. (Each writing is appended to the training text.)

3. Dasher is distributed under a Free Software License (no patents, no royalties, no license cost, and source code available for further developments) (FSF, 1991).

These characteristics suggest that Sign Language users would benefit from using dasher because reason 1 helps in sticking to a gesture approach for expression rather than switching to some sort of keyboard approach. The “diving” metaphor also allows users to have direct access to the whole Sign-Symbol-Sequence without having to fall back on symbol palettes or key-stroke combinations. The growing training text referred as the second reason improves prediction and enlarges its corpus. Such corpora based predictions are continuously adapting themselves to its users. There are reports that show that with prediction, dasher is even faster than a virtual keyboard (Ward and MacKay, 2002) or modified keyboard layouts. The third reason allows researchers, users and hackers to access the software source code for debugging, feature improvements or new developments.

In figure 1 we see an example of dasher zooming prediction. As the user selects a letter from the word been written (“Idea”, in the example), dasher zooms into the most likely letters to follow the previously selected ones (the context). The more a letter is likely to occur in the context³, the bigger its size gets. In figure 1, after have selected the sequence “I”, “d”, “e” the more likely letters are “n” (for “Identity”, “Identical”...), “a” (for “Idea”, “Ideal”, “Ideas”...) and so forth.

Dasher is available for use in dozens of languages. Using the application to write in any of those languages requires the user to set up an alphabet definition (that tells dasher which characters are valid and should be recognised in the chosen language) and a training text (a sort of corpus) written in that language.

3. Adapting Dasher for Sign language input

Agreeing with usefulness requirements as suggested by (Vettori et al., 2004), the choice of a sign language notation and an entry system to investigate the benefits of inference prediction in sign writing would have to fulfill the needs of both sign language researchers and users or, at least, try to do so. In addition to dasher (which is available or distributed within all major GNU/Linux distributions), we’ve chosen SignWriting as notational system because it is a broadly known notation after decades of usage and because it is used by some local deaf communities. Furthermore, SignWriting has an XML representation (SWML (Costa, 2009), (Costa and Dimuro, 2003)) that allows to interchange data with other SignWriting based tools.

We’ve assumed that LIBRAS⁴ has entropy comparable to written English, for simplicity reasons. An alphabet definition mapping SWMA2004 to Unicode glyphs was built and adaptations were made to the dasher source code to support

³According to the inference based on language model built up from the training text

⁴Brazilian Sign Language

this notation. A TrueType font to render signwriting symbols was compiled. A training text (in SWML), composed of children tales with restricted context and lexicon, was loaded as initial corpus.



Figure 2: “idea” in LIBRAS SignWriting notation

While for spoken languages the dasher diving canvas presents a linear character set (a to Z plus numbers and punctuation) as seen in figure 1, for signwriting use, the diving canvas was modified to match an alphabet definition to present a nested character set so that the user may dive into category, then group, choose the symbol in the next level and keep diving through variation, fill, and rotation to completely define the symbol to use.

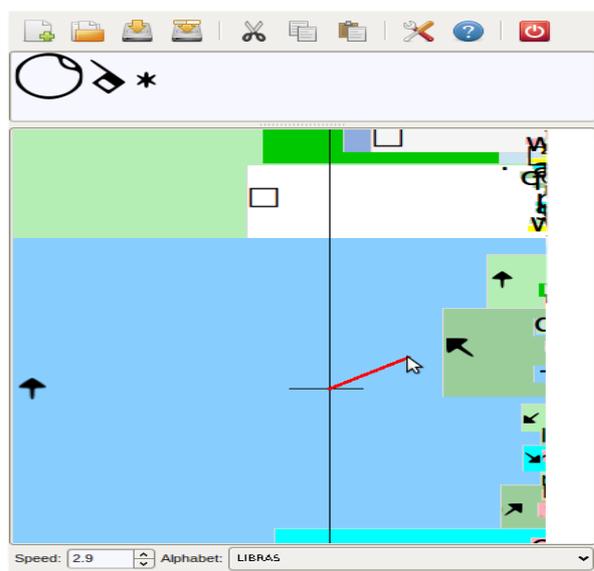


Figure 3: Writing “idea” in LIBRAS

4. Discussion, Conclusions and future work

A small group of occasional SignWriting users with varying signwriting skills were asked to perform writing tests. Results currently suggest a promising writing speed curve. Error rate results are inconclusive (as users perform more tests, some have increasing error rate while others have decreasing rates). Original output of dasher written texts, linear/left-to-right/top-down⁵, remains currently unchanged for signwriting in the present work. The task of recognising if written sign matches the intended one relies on user experience with signwriting and knowledge

⁵minor settings (as right-to-left) allowed

of LIBRAS. Error rates can be further investigated either by defining and running larger writing experiments or by defining an alternative to overcome the problem of linear output of dasher written texts. The alternatives may address issues by using matching algorithms. Through this framework, we suggest that using inference for Sign Language entry systems can speed up writing considerably. Improving text entry Efficiency for Sign Language may benefit not only research but also practitioners allowing them to access communication tools with more efficient writing and exchangeable format so they would be able, for instance, to use animated web instant messaging communication (Denardi et al., 2006).

The mentioned issues address several areas for future work, including address the spatial nature of SignWriting notation Vs. the linear writing offered by dasher; the unsettled SignSpelling that allows sign lexicographic ordering and searching should be studied in order to determine if it should be forced or corpora growth would lead to a long-term settling. The investigation can be reproduced using other notational system such as HamNoSys (Schmaling and Hanke, 2001), (Prillwitz and et al., 1987), ELiS (Estelita, 2008) or Stokoe (Stokoe, 1960) (Stokoe et al., 1965) or even other Sign Language notations.

5. Acknowledgements

This work would be possible without the inestimable support from SERPRO, ASL, UFRGS and prof. Dr. Paulo Blauth Menezes.

6. References

- A. C. R. Costa and G. P. Dimuro. 2003. SignWriting and SWML: Paving the way to sign language processing. In O. Streiter, editor, *Traitement Automatique des Langues de Signes, Workshop on Minority Languages*, Batz-sur-Mer, June 11-14.
- A. C. R. Costa. 2009. The SWML site. Located at: <http://swml.ucpel.tche.br>.
- Rubia M. Denardi, Paulo F. Blauth Menezes, and Antonio Carlos da Rocha Costa. 2006. An animator of gestures applied to the sign languages. In *2nd Workshop on the Representation and Processing of Sign Languages*, Genoa. ELRA, ELRA.
- Mariangela B. Estelita. 2008. *ELiS - Escrita das Linguas de Sinais: Proposta teorica e verificacao pratica*. Ph.D. thesis, UFSC - Universidade Federal de Santa Catarina. (Doutorado em Linguistica) Coordenao de Aperfeioamento de Pessoal de Nivel Superior.
- Free Software Foundation FSF. 1991. Gnu general public license, version 2. <http://www.gnu.org/licenses/old-licenses/gpl-2.0.html>.
- S. Prillwitz and et al., 1987. *HamNoSys. Hamburg Notation System for Sign Languages. An introduction*. Hamburg: Zentrum für Deutsche Gebärden Sprache.
- C. Schmaling and T. Hanke. 2001. HamNoSys 4.0. In T. Hanke, editor, *ViSiCAST Deliverable D5-1*, chapter Interface definitions. This chapter is available at <http://www.sign-lang.uni-hamburg.de/projekte/HamNoSys/HNS4.0/englisch/HNS4.pdf>.
- William C Stokoe, Dorothy C Casterline, and Carl G Croneberg. 1965. *A Dictionary of American Sign Language on Linguistic Principles*. , 2nd edition.
- W. Stokoe. 1960. Sign language structure: An outline of the visual communication systems of the American deaf. Buffalo, NY: Univ. of Buffalo. – E.E.U.U.
- Chiara Vettori, Oliver Streiter, and Judith Knapp. 2004. From CALL to Sign Language Processing: the design of e-LIS, an Electronic Bilingual Dictionary of Italian Sign Language and Italian. In Oliver Streiter and Antonio Carlos da Rocha Costa, editors, *Workshop on the Representation and Processing of Sign Languages*, 4th International Conference on Language Resources and Evaluation, LREC 2004, Lisboa, May.
- D. J. Ward and D. J. C. MacKay. 2002. Fast Hands-free writing by Gaze Direction. *Nature*, 418(6900):838.
- D.J. Ward, A.F. Blackwell, and D.J.C. MacKay. 2000. Dasher - A Data Entry Interface Using Continuous Gestures and Language Model. In *Proceedings of UIST2000*, page 129 137.

Towards semi-automatic annotations for video and audio corpora

S. Masneri¹, O. Schreer¹, D. Schneider², S. Tschöpel², R. Bardeli², S. Bordag³, E. Auer⁴, H. Sloetjes⁴ & P. Wittenburg⁴

¹Fraunhofer Heinrich-Hertz-Institute, Berlin, Germany

²Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, Sankt Augustin, Germany

³Max Planck Institute for Social Anthropology, Halle, Germany

⁴Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

E-mail: [stefano.masneri,oliver.schreer]@hhi.fraunhofer.de,
[daniel.schneider,rolf.bardeli,sebastian.tschöpel]@iais.fraunhofer.de, bordag@eth.mpg.de,
[eric.auer,han.sloetjes,peter.wittenburg]@mpi.nl

Abstract

AVATeCH (Advancing Video/Audio Technology in Humanities Research) is a project in which two Fraunhofer Institutes and two Max Planck Institutes collaborate in order to promote the development and application of technology for semi-automatic annotation of digital audio and video recordings. One of the aims of the AVATeCH project is to implement algorithms that allow for the automatic or semi-automatic creation of pre-annotations for the video corpora, hence reducing the time needed to perform the manual annotation task. Due to the huge size of the corpora, and the extreme variety of the video content, the algorithms developed need to be fast, efficient and robust. In this paper we will present some of the algorithms currently under development, the modifications applied in order to get them working with large video corpora and how the results of the annotations are stored, as well as how they can be integrated in ELAN annotation software.

1. Problem definition

In humanities research for psycho-linguistics, very large video corpora are available in order to investigate many different kinds of research topics such as relation between spoken language and gestures or preserving endangered languages. To evaluate the huge amount of video in the most efficient manner, meaningful annotations for the entire collection are required. These annotations should be of sufficient quality to both enable the user to gain an overview of the contents, as well as select and access important parts of the content quickly. One of the aims of the AVATeCH project is to implement algorithms that allow for the automatic or semi-automatic creation of pre-annotations for the video corpora, hence reducing the time needed to perform the manual annotation task.

Although the use of video analysis tools for automatic annotation has been a research field for many years, two aspects are different to common approaches. Due to the huge size of the corpora (approx. 30 TB), the algorithms need to be fast and efficient. Another important challenge is the diversity of the content. Though the most common case is that of persons being captured, almost any scenario can happen. Hence, very general, but also robust approaches need to be developed in order for the algorithms to be actually helpful in retrieving information out of the videos. A careful evaluation of robustness versus analysis quality needs to be taken into account. The almost arbitrary nature of the content does not allow an application of standard methods developed in the field of sign language recognition.

2. Video analysis algorithms

During the design of the algorithms for video analysis the

focus was mainly on the efficiency and the robustness of the solution. Efficient algorithms allow for faster automatic annotation, while robustness guarantees that meaningful annotations can be achieved for the majority of the content in the corpora. The design of algorithms that can perform well in many different scenarios (in the subset of corpora used for testing there are videos shoot in interview rooms, in restaurant, in small villages, in conference rooms, each of them with a different number of people represented) is the main guideline used to adapt existing solutions to the specific problem and to develop completely new approaches to the problem. The algorithms are designed to work in a fully automatic way, i.e. without the need for human interaction, in order to guarantee the possibility to use scripts to run the program on multiple files. The implementation is done using a highly modular structure, so that future algorithms can be seamlessly integrated in the current framework, using the results provided by the previous detectors.

2.1 Shot boundary detection

Shots consist of the video frames that have been continuously recorded with a single camera operation, and therefore represent the basic unit of a video. Since different shots refer to different camera operations, all of the detectors work on a shot basis. The tool developed in (Petersohn, 2004) was used as shot boundary detector, with few changes to the I/O interface. The shot boundary detector also retrieves the position of sub-shots inside of a shot. Sub-shots are defined as a sequence of consecutive frames showing one event or part thereof taken by a single camera act in one setting with only a small change in visual content. The detection of sub-shots has proved to be useful for the development of the other detectors.

A wrapper was added to the shot boundary detector, in order to make it work with all the types of video format used in the corpora and to provide a uniform, human readable and easy to parse XML output. Since all of the video in the corpora are unedited (and therefore there are neither fades nor wipes), only hard cuts are detected in order to improve the efficiency of the algorithm. The program runs faster than real-time, processing about 80 frames per second on standard definition videos.

2.2 Key-frames extraction

Because of the huge amount of data, it is extremely important to provide the user the possibility to efficiently browse the content of the videos. Since watching each video in the corpora would require hundreds of hours, the best option is to select an adequate number of images to represent all the content of a video.

The simplest (and somewhat obvious) approach is to extract an image every n frames, but it can lead to miss some important information or, on the other hand, extracting lots of almost identical images, when the scene is static and few changes happen. The new approach is to use the results of the shot-cut detection algorithm, extracting an image every time a new sub-shot is detected.

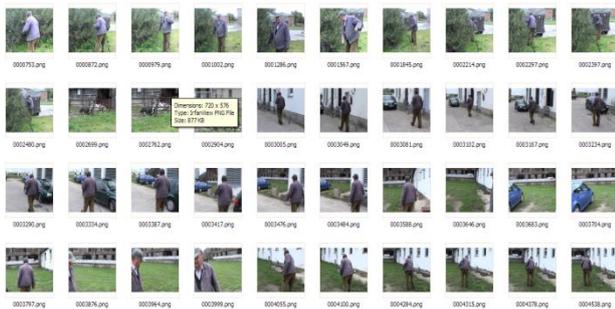


Figure 1: Results of key-frames extraction tool.

An additional image is extracted at the midpoint of a shot, to ensure that every shot in the video is represented by at least one image. With this approach, all the relevant information in a video is captured. Various options allow the user to decide the quality and the size of the output images. A typical use case (see **Figure 1** for an example) is to save the images as compressed thumbnails, allowing the user to grasp the content of the whole video with just a few glimpses.

The execution is usually five to ten times faster than real-time.

2.3 Global motion detection

The motion inside of a shot is another useful feature that can provide plenty of information to the user. An accurate motion analysis allows distinguishing between different types of video content. For example, an interview will have a static camera and a low amount of total movement inside the scene, while the video of a carnival will have lots of pans (i.e. camera motion) and motion inside of the scene. Further than that, an accurate motion analysis can provide helpful information for many other detectors.

The algorithm performs then a frame-based motion analysis and detects when global motion (pan or tilt) occurs inside of a shot. Our work is based on the *Hybrid Recursive Matching* algorithm (Atzpadin, Kauff & Schreer, 2004). For each frame in the shot, it extracts a motion map, representing the motion of a grid of pixels inside of the frame. The absolute value (i.e. the speed, calculated as L^2 norm) and the orientation are then computed, in order to obtain a vector field representing the total motion for that particular frame. An analysis of the motion map allows then to distinguish between camera motion and motion inside of the scene. This is particularly useful because in this way other than detecting camera motion, there is also the possibility to compensate motion when, as done in different detectors, moving objects inside of the scene needs to be tracked.

Once the motion for each frame in the shot has been analyzed, a post-processing step is performed, in order to reduce the fragmentation of the results. This step is necessary because the algorithm detects all kind of camera movements (even very short ones, e.g. when the person shooting the videos is adjusting the camera), while most of the time the user is interested only in longer camera movements. There is a set of parameters that can be tweaked by the user so that he can choose whether to have a very detailed but fragmented motion analysis or a coarser but less fragmented one. At the end of the algorithm a list of the camera motions occurring in each shot is obtained, with initial and final frame, as well as the direction of motion.

Working with standard definition video and taking motion vectors every 8x8 pixel, the program is able to process about 30 frames per second, slightly faster than real-time.

2.4 Skin colour estimation

In order to be able to successfully track objects inside of a scene the motion detector alone is not enough. To describe the object of interest other information needs to be extracted from the video. Since the user is typically interested in gesture annotation, it has been decided to focus on the detection and tracking of hands and heads. In order to perform the detection the first thing to do is *skin colour estimation*, that is to find the colour and luminance ranges that best represent the human skin. This is not an easy task, since for different videos skin is represented with different colours (skin colours vary depending on the person filmed, the luminance condition, the quality of the camera and other factors). Hence, the need is to develop an algorithm that allows accurate skin colour estimation without *a-priori* information.

It was chosen to work with the YUV colour space, instead of the more common RGB, for two reasons: the first one is that, under equal conditions, the U and V component tend to be similar for all kind of people filmed (e.g. the skin colour component of a Caucasian person and that of an Asiatic person are similar). The second one is that the biggest amount of information is contained in the luminance component Y, allowing us to use sub-sampled version for the U and V colour components, hence

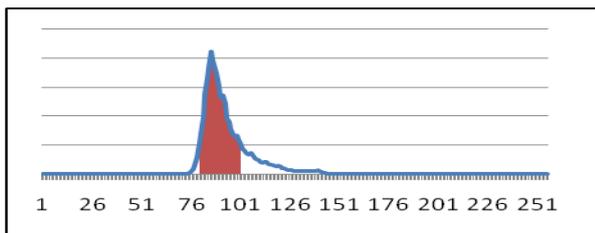


Figure 2: Histogram for U component. The range representing skin colour is highlighted

reducing the amount of memory needed and speeding up the computation.

The main goal of the skin colour estimation algorithm is then to narrow, for each one of the Y, U and V component, the range from [0...255] to the actual range corresponding to skin colour. An early, rough, estimation of the U and V ranges can be done by excluding the values that surely do not represent skin colour. Of course this narrowing operation is too approximate and must be further refined. To refine the U and V skin parameter estimation the results of the motion detector were used. In fact, it is assumed that the motion inside of a scene is most of the time due to the movement of a person, and therefore by identifying the motion inside a scene the presence of skin can be automatically identified as well. Of course there can be other moving objects in a scene, other than hands and heads, but these objects most of the time can be discarded since they are outside the colour ranges fixed before at the beginning of the motion analysis.

By repeating this procedure for each frame in the shot it is

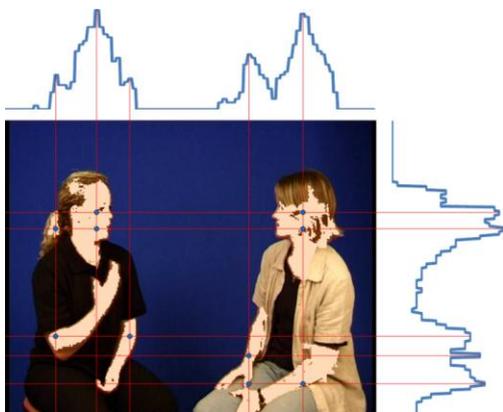


Figure 3: Detection of seed points

possible to keep track of all the pixels in the shot that most likely represent skin. By representing all the pixel values in a histogram the desired colour range can then be identified with a high degree of precision. **Figure 2** shows the histogram of the U values of the selected pixel for a shot: the highlighted range shows the interval that represents the skin range for this colour component.

Once the estimation of the U and V colour components is done, a new processing step is performed in order to estimate the luminance component. The execution time, for standard definition video, is comparable to real-time.

2.5 Hands and head detection and tracking

Once the colour and luminance ranges representing skin have been estimated accurately, the detection and tracking of hands and heads in the video is performed. The first step of the detection process involves the search of *seed* points where the hands and heads regions most likely occur. **Figure 3** shows how the seed points are selected: Histograms along the horizontal and vertical directions compute the number of pixels with luminance and colour values within the desired interval; the pixels where a maximum occur in both the directions are selected as seed points.

Starting from the seed points a region growing algorithm is applied, that selects all the points in the neighbourhood within the colour and luminance ranges found in the previous step. For each region found a different label is applied, allowing us to track the movement of different regions along the timeline. Each region found is then approximated with an ellipse, characterized by the position of the centre, the orientation and the length of the axes.

In **Figure 4** the result of the detection is shown: the pixels marked as skin are highlighted, and the ellipses approximating the region found are shown.

To discriminate between hands and heads a face detection algorithm (Viola & Jones, 2001) is applied and some basic geometric considerations (position of the heads, size of the region) are made in order to increase the robustness of the system.

The tracking is performed by considering the change of orientation and position of the ellipses along the timeline.

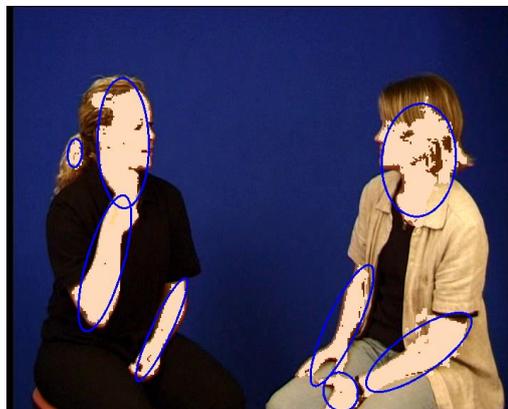


Figure 4: Hand and head tracking

The execution time depends upon the number of objects tracked, but is usually 3 to 5 times faster than real-time.

3. Annotation based on video analysis

Each one of the tools developed produces as output an xml file. The xml schema defining the syntax of the output files is the same for each program, in order to simplify the integration of the different tools and the addition of new ones.

For the shot boundary detection module the output xml contains a list of *shot* elements, each one of them has, as attributes, the start of the shot, the end of the shot and the position of the representative frame. Each shot element

can also have zero to many *subshot* children elements, specified by the relative representative frame. The subsequent annotation tools simply use the xml produced by the shot boundary detector as input, and add new children elements to each shot. The global motion detector adds *motion* elements, with attributes specifying the starting and ending frames, the type and the direction of motion. At last, the skin colour estimation tool adds one *segmentation parameters* element for each shot, with six attributes that specify the estimated intervals for the Y, U and V components.

The results of the hand and head tracking tool are not yet propagated to the output xml file, since various output options are currently being examined and no definitive choice has been made.

Below an example of a typical *shot* element, with all its child elements, is shown.

```
<shot start="0" end="501" keyFrame="250">
  <subshots>
    <additionalShot keyFrame="174"/>
    <additionalShot keyFrame="217"/>
    <additionalShot keyFrame="457"/>
  </subshots>
  <segmParams>124 133 134 81 10 16</segmParams>
  <motion>
    <cameraMotion direction="right" start="127" end="136"/>
    <cameraMotion direction="right" start="158" end="162"/>
    <cameraMotion direction="right" start="164" end="220"/>
  </motion>
</shot>
```

A *shot* element produced by the detectors

4. Integrating detectors in ELAN

The annotation software ELAN 3.6 introduced a Recognizer API for extending the program with pattern recognition components. The first implementation had limited functionality and only offered support for audio recognizers. The first attempts to integrate audio detectors proved to be particularly useful and generated a list of new wishes and requirements.

In the meantime, support for video recognizers has been added as well. To a large extent the interface for video recognizers follows the lines of that of audio recognizers, but the data structures for video differ slightly (no distinction between channels, provision for 2D area markers) and so does the user interface to interact with the recognizers. It is expected that more modifications of and extensions to the framework will prove necessary in the course of the project.

It needs to be noted that integration in ELAN is just one of the ways in which detectors will be made available; other solutions are developed as well. In the case of recognizers that are not deployed as extension of ELAN, the resulting XML can be imported into ELAN.

5. Future developments

Future developments point in two different directions. On one hand there is the possibility to improve the current detectors by increasing the richness of semantics. One

example is offered by the hand and head detection and tracking tool, for which it could be explored the possibility to map the tracked movement (e.g. rotation or movement in a certain direction) into actual hand gestures, in order to add to the framework gesture-recognition capabilities too.

On the other hand there is the development of tools that can extract other features (for example, a tool for tracking of other objects inside of the scene, for the segmentation of the video to allow distinguishing between background and foreground, for the creation of summaries of the videos).

6. Conclusion

A new framework for automatic annotation was developed. The framework is designed to be as robust and efficient as possible. It implements algorithms for fast browsing of the video corpora, motion detection, hand and head detection and tracking.

7. References

- Petersohn, C. (2004). *Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System*, TREC Video Retrieval Evaluation Online Proceedings, TRECVID.
- Petersohn C. (2007). *Sub-shots - basic units of video*. In EURASIP Conference Focused on Speech and Image Processing, Multimedia Communications and Services, Maribor, Slovenia.
- Boreczky J. & Rowe A. (1996). *Comparison of video shot boundary detection techniques*, Journal of Electronic Imaging, 5(2), 122-128.
- Atzpadin, N., Kauff, N. & Schreer, O. (2004). *Stereo Analysis by Hybrid Recursive Matching for Real-Time Immersive Video Conferencing*, Trans. on Circuits and Systems for Video Technology, Special Issue on Immersive Telecommunications, Vol.14, No.3, pp. 321-334.
- Askar, S., Kondratyuk, Y., Elazouzi, K., Kauff P. & Schreer, O. (2004). *Vision-Based Skin-Colour Segmentation of Moving Hands for Real-Time Applications*, Proc. of 1st European Conf. on Visual Media Production (CVMP 2004), London, United Kingdom.
- Lausberg, H., & Sloetjes, H. (2009). *Coding gestural behavior with the NEUROGES-ELAN system*. Behavior Research Methods, Instruments, & Computers, 41(3), 841-84.
- Crasborn, O., Sloetjes, H., Auer, E., & Wittenburg, P. (2006). *Combining video and numeric data in the analysis of sign languages with the ELAN annotation software*. In C. Vettori (Ed.), Proceedings of the 2nd Workshop on the Representation and Processing of Sign languages: Lexicographic matters and didactic scenarios (pp. 82-87). Paris: ELRA.
- Viola P. & Jones, M. (2001). *Robust real-time object detection*. Second International Workshop on Statistical and Computational Theories of Vision.

Dealing with Sign Language Morphemes in Statistical Machine Translation

Guillem Massó, Toni Badia

Grup de Lingüística Computacional (GLiCom)
Universitat Pompeu Fabra, Roc Boronat 138, Barcelona, Spain
E-mail: guillem.masso@upf.edu, toni.badia@upf.edu

Abstract

The aim of this research is to establish the role of linguistic information in data-scarce statistical machine translation for sign languages using freely available tools. The main challenge in statistical machine translation is the scarcity of suitable data, and this problem becomes more pronounced in sign languages. The available corpora are small, usually not domain-specific, and their annotation conventions can vary considerably. Elaborating our own corpus is a very time-consuming task and the amount of data that we can obtain is even more reduced. Under these conditions, morpho-syntactic information helps to improve statistical machine translation results, but there are not linguistic processing tools for sign languages. We have managed to improve translations from Catalan to Catalan Sign Language by using factored models in an open source translation system with basic linguistic information such as the lemma or an annotation tier tag. Furthermore, this allows us to deal with sign language morphemes in a more systematic way.

1. Introduction

Nowadays, there is an increasing interest in data-driven approaches in machine translation (DDMT), either statistical (SMT) or example-based (EBMT): their development is less time-consuming and they are more scalable than rule-based approaches, although a considerable amount of data is required to create bilingual corpora. So, the main challenge is to set up a suitable parallel corpus large enough. Problems in SMT due to scarce resources, which are endemic in sign languages (SLs), have also been detected in oral languages (OLs). One of the suggested solutions to improve translation results is to use morpho-syntactic information (Nießen and Ney, 2004). This is a good solution if there are linguistic processing tools for the analysed languages: once more, SLs are at a disadvantage. However, these tools can be used for the OL corpus part, and other alternatives must be found for the SL analysis. In this work, we propose two solutions: the use of plain glosses as lemmas of inflected forms, and the use of annotation tier names as tags in a more syntactical approach. This linguistic information is integrated at the word level using factored models of Moses, an open source SMT system (Koehn et al., 2007).

The remainder of the paper is organized as follows. In section 2, we give a brief overview of related work in DDMT and the use of morpho-syntactic information. In section 3, we present our parallel corpus for the language pair Catalan and Catalan Sign Language (LSC). Section 4 describes the experiments carried out and section 5 discusses the results and evaluation. Finally, section 6 outlines the main conclusions of the work.

2. Related work

As for sign language MT (SLMT), and as far as we are aware, there are four main research groups working on DDMT. Stein, Bungeroth and Ney (2006) use a

phrase-based SMT system for the language pair German and German SL (DGS). The SL corpus is annotated with glosses, including all important grammar features. Their research is focused on morpho-syntactic pre and post-processing enhancement. In the pre-processing step, German is analysed by a parser, and part-of-speech (POS) information is used to transform nouns into stem forms, split compound words and delete German POS not used in DGS. In the post-processing step, marked positions of discourse entities are added from a database. Some deleted information about emphasis and comparative degree is added as well. Therefore, morpho-syntactic information is not used during the translation process.

The research of Morrissey and Way (2007) focuses on EBMT. The ATIS corpus (Bungeroth et al., 2008) was translated from English to Irish SL (ISL) to be used as data set. The SL data are annotated with glosses but without non-manual or phonetic feature detail, and no morpho-syntactic information is used.

The two aforementioned groups have collaborated in Stein et al. (2007) and Morrissey et al. (2007) to translate from SL to OL with SL recognition. Although their research does not focus on morpho-syntactic improvements in SLMT, some interesting issues are raised. The main one concerns the handicap of lacking SL parsers, since morpho-syntactic information usually reduces errors. However, the authors consider that adding features such as the hand tracking position in pointing signs is comparable to adding POS information. They also suggest that “other features are likely to improve the error rates as well and should be investigated further” (Stein et al., 2007).

Su and Wu (2009) go beyond and use a treebank, a bilingual dictionary and a translation memory to convert the Chinese syntactic structure with thematic role information into the corresponding structure in Taiwanese SL. Thematic roles also allow them to deal with

agreement verbs by identifying verb arguments and providing movement directions. However, the authors highlight that the proposed system hardly deals with non-manual features, although this issue would be the next step in their research.

San-Segundo et al. (2008) work on speech recognition and MT from Spanish to Spanish SL (LSE). They compare a rule-based MT system with a SMT system. The rule-based system obtains the best results: on one hand, the restricted domain (a service for renewing identity cards) makes it possible to develop a complete set of rules with reasonable efforts, and on the other hand, the statistical system cannot be trained properly due to the reduced amount of data. They also collaborated with the Aachen group (D'Haro et al., 2008) to improve the sign language model using information retrieval from the Web.

3. Corpus

Nowadays, there is not any available corpus in LSC which could be used for MT, so we have created a small corpus on the weather report domain. This is a restricted domain with a limited vocabulary that allows us to obtain reasonable results with scarce resources. The original Catalan texts were retrieved from the Catalan Weather Service website (Servei Meteorològic de Catalunya¹) and translated by a native Deaf signer. Catalan sentences were analysed with the freely accessible tagger CatCG² (Alsina et al., 2002) to obtain lemmas and POS, and they were manually revised. The recorded LSC sentences were annotated with iLex (Hanke & Storz, 2008), which allows a greater control over the annotation process thanks to its lexical database.

We were especially interested in morphemes containing adverbial and aspectual information. In order to systematically annotate these linguistic features, the gloss tier contains plain glosses, which we will consider lemmas, and there are separated tiers for mouth morphemes and for movement morphemes. Regarding annotation, the currently available guidelines (Neidle, 2002; Nonhebel, Crasborn & van der Kooij, 2004) do not offer a suitable description for the analysed LSC morphemes, so specific tags have been created. However, the important thing is not the tag assigned, but the fact that morphemes are individualised and classified.

We made two sets from the annotation files. Both sets have added factors with linguistic information, but they differ in SL morphemes representation. In set 1, morphemes are attached to glosses and the lemma is a factor, represented by the plain gloss. In set 2, morphemes are independent tokens and the added factor is the annotation tier name. It can be seen in the next example, where the vertical bar separates factors, *ct* stands for the mouth morpheme *cheeks puffed and tense*, and *f* stand for

the movement morpheme *fast movement*. This example means 'heavy rain':

Set 1: RAIN:ct:f|RAIN

Set 2: RAIN|gloss ct|mouth f|movement

Statistics of the bilingual corpus are shown in Table 1. Notice that there are not lemmas in set 2 because there is not form variation.

		Catalan	LSC (Set 1)	LSC (Set 2)
Training	Sentences	153		
	Running words	1967	1520	1930
	Vocabulary	282	220	182
	Lemmas	241	162	n/a
	Singleton words	87	77	50
	Singleton lemmas	66	46	n/a
Test	Sentences	46		
	Running words	449	376	479
	Vocabulary	164	130	116
	Lemmas	146	102	n/a
	Singleton words	88	64	45
	Singleton lemmas	70	41	n/a
	OOV words	10	5	2
	OOV lemmas	7	2	n/a

Table 1: Statistics of the bilingual corpus with two annotation sets for Catalan Sign Language (LSC).

4. Experiments

The system used is Moses (Koehn et al., 2007), an open source toolkit for SMT. Moses relies on SRILM (Stolcke, 2002) to create language models (LM) of the target language, and on GIZA++ (Och & Ney, 2003) for the alignment process. This system enables the integration of additional information at the word level using factored models. As for the OL, we use the lemma and the POS as added factors. As for the SL, the added factor is the lemma in set 1, and the annotation tier in set 2, as mentioned in the previous section.

In previous tests, we noticed that using a smaller training set plus a development set to tune the translation models gives worse results than using a bigger training set without tuning, probably due to the small amount of data. In the end we decided to train and tune the system with the whole training set in order to optimize the results. The LM was also improved by considering all the available sentences of the training and test sets. It is important to highlight that the system creates one LM for each factor of the target language. The built LMs are based on tri-grams.

Given that the aim of these experiments is to evaluate the role of linguistic information, the factors of source and target languages are combined in different ways. As for the source language, translations are from: form, lemma, form + lemma, lemma + POS, form + lemma + POS. As for the target language, translations are to: form, form +

¹http://www.meteo.cat/mediamb_xemec/servmet/index.html

²<http://www.glicom.upf.edu/projectes/catcg>

factor (lemma or annotation tier). Altogether, there are ten translations per set.

5. Results

5.1 Machine evaluation

All the translations have been evaluated with the NIST and BLEU metrics, as can be seen in Table 2. The most relevant fact is that translations with an added factor for the target language (set *b*) are considerably better than translations to only the target form (set *a*). As for the source language factors, it is not always clear that they can improve the translation. Differences between set 1 and set 2 depend on factors as well, and the two metrics are not always coherent.

		Set 1		Set 2	
		NIST	BLEU	NIST	BLEU
Set <i>a</i>	F→F	5.0682	0.4427	5.2071	0.3967
	L→F	5.6373	0.4958	5.3476	0.4307
	F+L→F	5.5198	0.4700	5.2515	0.3908
	L+POS→F	5.7826	0.5059	5.2255	0.3939
	F+L+POS→F	5.4497	0.4596	5.4658	0.4378
Set <i>b</i>	F→F+AF	6.8178	0.6294	6.3251	0.6951
	L→F+AF	6.6842	0.6373	6.3809	0.7245
	F+L→F+AF	6.8968	0.6271	6.1389	0.6783
	L+POS→F+AF	6.5717	0.6172	6.4224	0.7111
	F+L+POS→F+AF	6.9004	0.6234	6.4110	0.7143

Table 2: Machine evaluation results.
(F = form, L = lemma, AF = added factor)

In subset *1a*, the worst results are for translations from the surface form, and the maximum improvement is of 0.7144 in NIST and 0.0632 in BLEU by using the lemma and the POS. The second best score is for translations from the lemma. Nevertheless, if the three factors are used (form + lemma + POS), the second worst result is obtained. On the other hand, in subset *2a*, the latter combination is the best, and the second best score is again for the lemma. The differences among the other three options are rather low. The score variability in subset *2a* is of 0.2587 in NIST and 0.0470 in BLEU. In general, scores are better in subset *1a* than in subset *2a*.

In subset *1b*, the worst results are for translations from the lemma and the POS. In the other cases, the metrics are not coherent. In NIST, the best scores are for (in this order): form + lemma + POS, form + lemma, form, lemma. In BLEU, the order is inverted. The score variability is of 0.3287 in NIST and 0.0201 in BLEU. In subset *2b*, translations from form + lemma obtain the worst results, followed by translations from only the form. The best scores in NIST are for lemma + POS, form + lemma + POS and lemma. In BLEU, for lemma, form + lemma + POS and lemma + POS. The score variability is of 0.2835 in NIST and 0.0462 in BLEU. Within the set *b*, NIST

scores are higher in subset *1b*, while BLEU scores are higher in subset *2b*.

While the maximal improvement by combining source factors has been of 0.7144 in NIST and 0.0632 in BLEU, the improvement by adding one target factor has been of 0.7891-1.7496 in NIST and 0.1113-0.3172 in BLEU. This is probably due to the fact that the system has two related LMs, which improves the quality of the target sentences, although the LM had already been optimized. Considering these results, the improvement of the LM seems to be more important than the improvement of the translation model. In addition, it is difficult to find clear patterns for the role of source factors in the translation process.

5.2 Human evaluation

Unfortunately, it was not possible to conduct a human evaluation by native Deaf signers. Nevertheless, it is interesting to analyse some translation results in order to clarify the role of source factors and the differences between set 1 and set 2. Four translations were chosen: form → form, form + lemma + POS → form, form → form + factor, form + lemma + POS → form + factor. We evaluated the sentences from 1 (wrong) to 5 (correct) and we noticed that 27 sentences had been correctly translated in all the cases. These sentences fulfil two conditions: they have been seen in the training set and their length is equal or lower than 10 words. As their translation difficulty is low, we will analyse the other 19 sentences, 3 of which are seen sentences longer than 10 words and 16 are not seen sentences. The number of sentences for each score and the average per sentence are shown in Table 3.

		Score					Average
		5	4	3	2	1	
Set 1	F → F	3	4	6	5	1	3.11
	F+L+POS → F	5	1	8	5	0	3.32
	F → F+AF	4	2	8	5	0	3.26
	F+L+POS → F+AF	5	4	6	4	0	3.53
Set 2	F → F	3	0	4	9	3	2.53
	F+L+POS → F	1	5	5	7	1	2.89
	F → F+AF	1	1	6	9	2	2.47
	F+L+POS → F+AF	3	4	7	4	1	3.21

Table 3: Human evaluation results by number of sentences for each score.

(F = form, L = lemma, AF = added factor)

Concerning the differences between the two sets, the scores of set 1 are clearly higher than the corresponding ones of set 2. We have noticed that set 2 has more syntactic errors due to incorrect positions assigned to morphemes and to wrong gloss-morpheme combinations. As for the factors considered, the best results are obtained with all of the factors of both languages. It is important to highlight that the improvement by adding the source

factors is higher than that by adding the target factor, contrary to what machine evaluation shows.

6. Conclusions

Although a complete human evaluation by native Deaf signers would be necessary, we can assert that factored models with linguistic information for both source and target languages improve the results of statistical SLMT. Regarding the SL, complex morpho-syntactic analyses are not indispensable, but simple information from annotation files can be used in an efficient way. Furthermore, this allows us to deal with SL morphemes, which are usually ignored in SLMT. The analysis of the results shows that the best solution of the two proposals is to attach morphemes to glosses and to use plain glosses as lemmas, which are used as added factors. The other solution, considering morphemes as independent tokens, can generate additional syntactic errors.

7. Acknowledgements

This research was supported by the Commission for Universities and Research of the Department of Innovation, Universities and Enterprise of the Catalan Government (Generalitat de Catalunya) and by the European Social Fund.

8. References

- Alsina, À., Badia, T., Boleda, G., Bott, S., Gil, À., Quixal, M., Valentín, O. (2002). CATCG: A General Purpose Parsing Tool Applied. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas de Gran Canaria, Spain, 3, pp. 1130—1134.
- Bungeroth, J., Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A., van Zijl, L. (2008). The ATIS Sign Language Corpus. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, pp. 2943—2946.
- D'Haro, L.F., San-Segundo, R., Córdoba, R., Bungeroth, J., Stein, D., Ney, H., (2008). Language Model Adaptation for a Speech to Sign Language Translation System using Web Frequencies and a MAP Framework. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH 2008)*, Brisbane, Australia, pp. 2199—2202.
- Hanke, T., Storz, J. (2008). iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora (LREC 2008)*, Marrakech, Morocco, pp. 64—66.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, pp. 177—180.
- Morrissey, S., Way, A. (2007). Joining Hands: Developing a Sign Language Machine Translation System with and for the Deaf Community. In *Proceedings of the Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments: Assistive Technology for all Ages (CVHI 2007)*, Granada, Spain.
- Morrissey, S., Way, A., Stein, D., Bungeroth, J., Ney, H. (2007). Combining Data-Driven MT Systems for Improved Sign Language Translation. In *Proceedings of the Machine Translation Summit XI (MT'07)*, Copenhagen, Denmark, pp. 329—336.
- Neidle, C. (2002). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. *American Sign Language Linguistic Research Project Reports*, 11.
- Nießen, S., Ney, H. (2004). Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. *Computational Linguistics*, 30(2), pp. 181—204.
- Nonhebel, A., Crasborn, O., van der Kooij, E. (2004). *Sign language transcription conventions for the ECHO Project*. Radboud University Nijmegen.
- Och, F.J., Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19—51.
- San-Segundo, R., Barra, R., Córdoba, R., D'Haro, L.F., Fernández, F., Ferreiros, J., Lucas, J.M., Macías-Guarasa, J., Montero, J.M., Pardo, J.M. (2008). Speech to sign language translation system for Spanish. *Speech Communication*, 50(11-12), pp. 1009—1020.
- Stein, D., Bungeroth, J., Ney, H. (2006). Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT 2006)*, Oslo, Norway, pp. 169—177.
- Stein, D., Dreuw, P., Ney, H., Morrissey, S., Way, A. (2007). Hand in Hand: Automatic Sign Language to Speech Translation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden, pp. 214—220.
- Stolcke, A. (2002). SRILM – An Extensible Language Model Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, Denver, USA, vol. 2, pp. 901—904.
- Su, H.-Y., Wu, C.-H. (2009). Improving Structural Statistical Machine Translation for Sign Language with Small Corpus Using Thematic Role Templates as Translation Memory. *IEEE Transactions on Audio, Speech and Language Processing*, 17(7), pp. 1305—1315.

Elicitation Tasks and Materials designed for Dicta-Sign’s Multi-lingual Corpus

Silke Matthes¹, Thomas Hanke¹, Jakob Storz¹, Eleni Efthimiou², Nassia Dimou², Panagiotis Karioris², Annelies Braffort³, Annick Choisier³, Julia Pelhate⁴, Eva Safar⁵

¹Institute of German Sign Language and Communication of the Deaf, University of Hamburg, ²Institute for Language and Speech Processing / “Athena” R.C., ³LIMSI/CNRS, ⁴WebSourd, ⁵University of East Anglia
 {silke.matthes,thomas.hanke,jakob.storz}@sign-lang.uni-hamburg.de, {eleni_e,ndimou,pkarior}@ilsp.gr,
 {annelies.braffort,annick.choisier}@limsi.fr, julia.pelhate@websourd.org, E.Safar@uea.ac.uk

Abstract

This paper presents elicitation tasks and materials designed for the Dicta-Sign project. Within the framework of the project, sign language corpora are being compiled for four European sign languages. The aim for the data collection was to achieve as high a level of naturalness as can be achieved with semi-spontaneous utterances under lab conditions. Therefore, informants were filmed in pairs interacting with each other. With respect to parallelisability, elicitation tasks had to be designed that result in semantically close answers without predetermining the choice of vocabulary and grammar. The tasks developed for this purpose cover different interaction formats ranging from monologues to sequences of very short turns, also with different levels of predictability. The materials designed as well as experiences gained adjusting and using the material for Dicta-Sign’s different target languages are illustrated in this paper.

1. Introduction

The Dicta-Sign project, which started in January 2009, has the major objective to enable communication between Deaf individuals by promoting the development of natural human computer interfaces for Deaf users. It will research and develop recognition and synthesis systems for sign languages at a level of detail necessary for recognising and generating authentic signing. Research outcomes will be integrated in three laboratory prototypes:

- A Search-by-Example Interface to a Multilingual Lexical Database
- A domain-specific Sign-Language-to-Sign-Language Translator
- A Sign-Wiki (a signing avatar presenting the information).

Dicta-Sign deals with four European sign languages: British Sign Language (BSL), German Sign Language (DGS), Greek Sign Language (GSL) and French Sign Language (LSF). As one of the first steps, sign language video corpora have to be compiled for all of the target languages consisting of about 5 hours of annotated video per language. In the currently ongoing data collection Deaf informants are filmed in pairs, with each recording session lasting about two hours. Elicitation tasks and materials were developed specifically for the project’s purpose, aiming at building corpora parallelised as much as possible.

2. Corpus Content

Parallel corpus collection for sign languages has so far been undertaken only in minimal sizes or for spoken language simultaneously interpreted into several sign languages, but not for semi-spontaneous signing by native signers. The “oral” nature of sign language as well as the risk of influences from written majority languages

complicate the collection of parallel corpora. In fact, corpus planning needs to balance between naturalness of the data to be collected and the degree of parallelisability of the data across languages. The decision taken for Dicta-Sign was to aim at as high a level of naturalness as can be achieved with semi-spontaneous utterances under lab conditions. One key point here was to film Deaf informants in pairs, interacting with each other. With respect to parallelisability, elicitation tasks had to be designed that result in semantically close answers without predetermining the choice of vocabulary and grammar.

The domain selected for Dicta-Sign is travel across Europe. This is a domain of interest for Deaf people, and it combines general knowledge with personal experiences. On the sign language side, this domain offers great potential to elicit signing space construction in various dimensions for all of the target languages, but also allows for elicitation formats coming close to the goal of a parallel corpus.

The elicitation tasks are targeted towards a session length of about two hours. With a target number of sessions of eight (i.e. sixteen informants) for each target language, this will result in video material well beyond the target size of the corpus (i.e. 5 hours from 10 different signers per language). While it is highly unlikely that all recordings can be annotated later in the project, this approach also leaves room to exclude parts of the corpus data if needed.¹

¹ This might become necessary for a number of reasons, e.g. one of the informants revealed very private personal experiences that he or she later prefers to be excluded from the corpus to become publicly available, or it turns out that an informant’s language fluency is not as expected. Also, the size leaves more flexibility in choosing data regarding the parallelisability of the corpus.

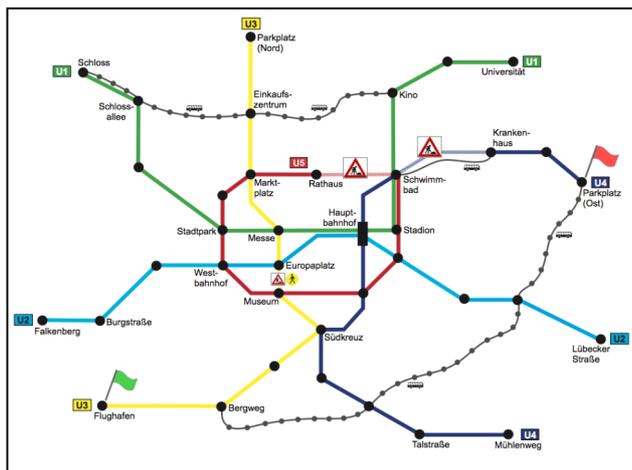


Figure 2: Transportation map (German material for informants A and B)

3.2 Description of Places and Activities

For the following tasks pictures³ are provided to evoke language use in a less restricted way than in the tasks previously described. At the same time a high percentage of vocabulary comparable across the different languages can be expected.

3.2.1 Travel Agency

The imagined setting for this task is a travel agency. The role of the one informant is that of a staff member at the travel agency, while the other takes over the role of a customer who wishes to go on holiday but does not know yet where to go to. The first informant is asked to suggest two different destinations to the customer by describing/advertising these places. Afterwards the second informant is asked to explain briefly which of the destinations she/he prefers and why.

The destinations to be described are predetermined by the elicitation material and vary from session to session in order to cover a wider range of vocabulary. For each session this is one capital city (Paris, London, Athens or Berlin) and a more general place (either “at the beach” or “in the mountains”). For each destination a range of pictures are shown to the informant meant to provide ideas of what to talk about. Included are well-known tourist places, other places of interest (e.g. museums), certain characteristics of this place (e.g. different styles of houses/living), places specifically of interest to Deaf people (e.g. Deaf Theatre), leisure activities, etc. The pictures of a certain destination are presented one after the other (each shown for 2 sec), which prevents the informant from concentrating on each detail of a picture. At the end of the presentation a collage of the pictures is shown which remains displayed throughout the task.

The selection of pictures included in the material as well as the destinations varying from session to session ensure

that a wide range of domain specific vocabulary is covered.



Figure 3: Travel Agency: Paris (material for all languages)

3.2.2 At the Airport

The topic of this task is the situation at an airport and the procedures taking place when travelling by plane. The informant is asked to describe the situation as if the other informant has never travelled by plane before. Pictures displaying different aspects as checking in, security issues, boarding, baggage claim and passport control are shown to the informant in chronological order and displayed as a collage at the end. Again the pictures are not to be described in detail but meant to provide ideas of what to talk about.

Mainly monologue data is expected from this task, but the design of the task also allows for involvement of the second person adding to the other person’s description.

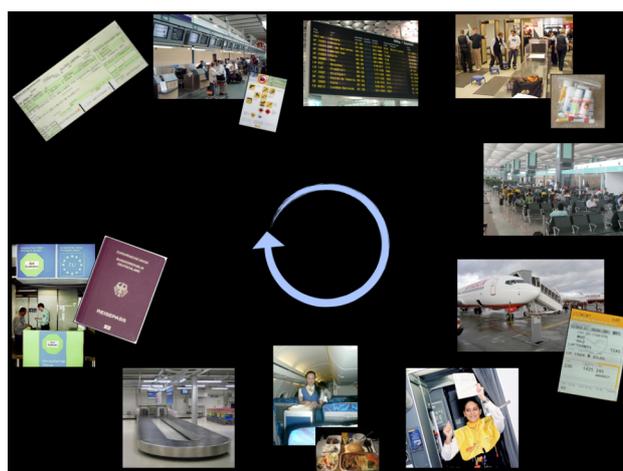


Figure 4: At the airport (German material)

³ Pictures used for this and all collages in other tasks were published under Creative Commons licenses (URLs available upon request).

3.3 Discussion and Negotiation

The following section deals with collaborative tasks. Taking the risk of receiving utterances that are less comparable than those of other tasks (not only across languages but also for the individual informants), these tasks are aiming at language data coming closer to a natural interaction.

3.3.1 Planning a Holiday

In this task the informants are instructed to plan a holiday together. They are asked to negotiate the destination, the time period for the holiday and the means of transport. They should also take into account further aspects relevant for their decision. As a basis of their discussion a picture card is shown to each of the informants displaying flags of several countries/regions, different means of transport, as well as a calendar where certain time periods are blocked (only the calendar is different for the two informants). Additionally certain aspects are shown that should also be taken into account (weather, temperature, costs).

A high amount of interaction is expected from this task, also providing domain-specific vocabulary that is not covered by the other tasks (especially dates, time periods, etc.).

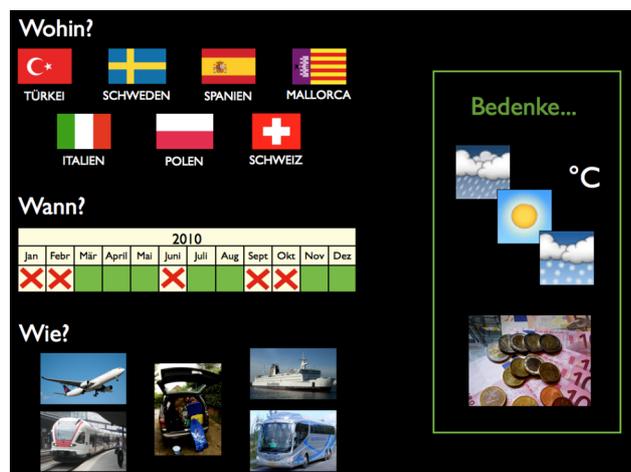


Figure 5: Planning a holiday (German material for informant A)

3.3.2 Travel Then & Now

The informants are asked to discuss how travelling has changed over time. The task is not restricted to a specific content, however pictures are presented to both informants in order to provide ideas. They show different aspects, e.g. means of transport and distances, passport control, money, up to booking on the internet and low-cost airlines. The presentation of the stimuli is similar to the one described above (see 3.2).

Depending on the individual informants the stimuli might lead to a different degree of interaction. The aim is

to provoke a discussion between the informants, possibly enhanced by a narration of personal experiences.⁴

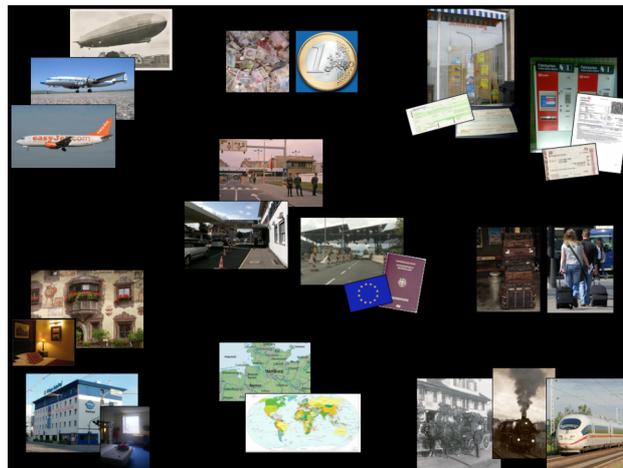


Figure 6: Travel then & now (German material)

3.4 Narration

The tasks described in this chapter are narrative tasks with varying degree of content predetermination. Whilst for 'Expectation & Reality' the setting of the story is given but not the exact content, the other two tasks ask for renarration of a given story.

3.4.1 Expectation & Reality

The informants are asked to tell short stories based on picture cards showing two opposed occurrences of a certain situation (somebody's expectations and the actual situation). The following situations are included in the task (three picture cards for each informant plus one example):

- Skiing holiday / no snow (example)
- Comfortable hotel room / tiny room with small bed
- Summer holiday / flight cancelled
- Visit to a nice museum/ overcrowded museum
- Plenty of food in a restaurant / plate with little food
- Garden party with BBQ / bad weather
- Sunset at the beach / traffic jam, arriving in the dark

The informants are free to tell a true story (where something similar has happened to themselves) or make one up. An example is given during the task explanation in order to show what kind of story they are asked for (i.e. first-person narration, adding information and developing a story line, length of the story). While the content of the elicited stories is less predictable, this task aims at as high a naturalness of the data as possible, meanwhile providing vocabulary related to the target domain.

⁴ The moderator is asked to encourage the informants if needed.

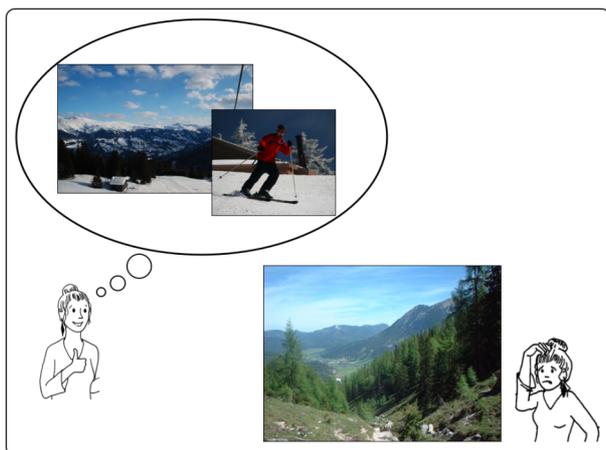


Figure 7: Expectation and reality (example for all languages)

3.4.2 Picture Story

A picture story by Quino (Lavado 1991) is used for this task, in which a woman explains to a tourist how to get to a certain restaurant (including walking, taking a taxi and a plane). One of the informants is asked to look at the story picture by picture and tell it to the other informant afterwards.

As the content of the story is given, the task is expected to provoke monologue data relatively similar by content for all target sign languages.

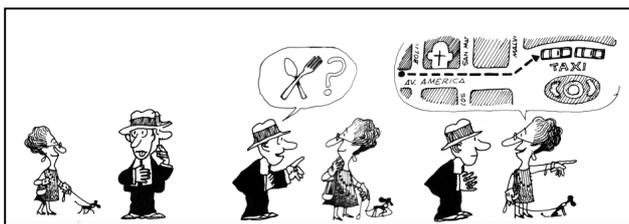


Figure 8: Beginning of the picture story by Quino

3.4.3 Retelling of a Signed Story

The informant being the recipient in the previous task is now asked to watch a video clip of a signed story and renarrate it afterwards. The fictive story is told by a Deaf person and deals with what happened during his last holiday: His travel group arrived late at the hotel, and after a drink at the bar he went straight to bed and slept all night. The next morning the others tell him that the fire alarm went off during the night and how they tried to wake him up (clip length approx. 2min).

The story has originally been produced in DGS alongside with a written English translation and was then translated in each of the other target sign languages. Due to the given content of the story and the sign language input the elicited data is expected to be comparable across the individual informants as well as the different target languages. A high amount of sign language characteristic features (e.g. Constructed Action, nonmanuals) can be expected from this task.

3.5 Denomination

In addition to the other tasks it was decided to elicit a number of signs in isolation in order to ensure that the corpus will contain certain vocabulary relevant to the target domain. Phrases however are not included in the elicitation due to the risk of major influence from spoken language.

3.5.1 Isolated Signs

Single pictures and, where needed for clarification, written words are used as stimuli. The informants are asked in turns to give isolated signs for the concepts shown. They are also encouraged to add to the other person's answers whenever they know a different sign for a certain concept.

The task covers the following areas: dates (days of the week, months and numbers), vehicles, countries of the EU, weather conditions.

3.6 Task Explanations for the Informants

A Deaf moderator is present during the whole elicitation session ensuring a smooth procedure, providing support for the informants and being responsible for the time management.⁵ However, consistently explaining the tasks to the informants is a complex issue that cannot easily be done offhand during the elicitation. The phrasing needs to be planned carefully and Deaf culture-specific aspects regarding the text structure need to be considered. It was therefore decided to film all the explanations beforehand and show these clips to the informants prior to each task. This still leaves a lot of responsibility to the moderator leading the elicitation but ensures that no information is left out and that each informant gets exactly the same explanation (especially across different languages and with varying moderators).

3.7 Procedure for the Elicitation Sessions

The tasks described above are to be arranged in a way assuring a balance with respect to the activity of both informants in a session. Switching roles between the topics was arranged as shown in the timetable below. The estimated duration for each task given in the timetable includes the task explanations given to the informants (aiming at a total session length of about two hours).

Each session starts with a *warm-up task*, where the informants are introduced to the domain of the elicitation and are led into a short conversation about their own travel habits. A short *break* is planned for between tasks 5 and 6, and the session is concluded with a slot for the informants' feedback. Additionally an *extra task* has been planned for in case the estimated time for a session is not fully used. For this task, no material is shown but the informants are asked to tell a personal travel experience (e.g. their best or worst holiday ever).

⁵ For the moderator's role see Hanke et al.: DGS Corpus & Dicta-Sign: The Hamburg studio setup, this volume.

No.	Task	Informant A	Informant B	Estimated dur. (min)
0	<i>Warm up</i>	<i>conversation</i>		5
1	Public transportation	explanation	explanation	10
2	Travel agency	description	(<i>short answer</i>)	11
3	Planning a holiday	negotiation		7
4	At the airport	-	description	5
5	City Map	explanation	(<i>follows</i>)	9
	<i>Break</i>			5
6	Expectation & reality	narration	narration	12
7	Travel then & now	discussion		11
8	Retell a story	narration (<i>signed story</i>)	narration (<i>picture story</i>)	10
9	Isolated signs	denomination	denomination	12
10	<i>Extra task:</i> Pers. experience	narration	narration	10
11	<i>Feedback</i>	<i>comment</i>	<i>comment</i>	6

Table 1: Procedure for elicitation sessions

3.8 Material Adaptations

In planning parallelised corpora of different languages, also cultural differences as well as language-dependent issues have to be taken into account. The material was therefore designed in a way that only adjustments are needed that are easy to realise and do not change the character of a tasks.⁶

Obviously adjustments are needed for tasks that include written language. Mostly the words can easily be translated; a version of the materials just including the drawings and pictures can be used for the adaptation, where only the words have to be added. The only exception is the ‘Public transportation’ task: While most of the stations are named after locations (e.g. town hall or hospital), some are typical street names that can not be translated directly (e.g. Kings Road, Green Lane) but were chosen to elicit utterances that include signs as “road” or “place”. Additionally for the ‘City map’ material icons are used alongside with the written words for an easy comprehension. These need to be changed according to the usage in each country (e.g. pharmacy). Several tasks rely on pictures as stimuli that can mostly be used across the different target languages. Some pictures however are country or language specific and need to be replaced (e.g. passport, train ticket, typical kind of hotel).⁷ Additionally, the task evoking isolated signs allows for pictures to be added in case a specific

⁶ This holds for Dicta-Sign’s target languages and presumably for other sign languages in Europe and beyond.

⁷ Most of the changes are needed for the tasks ‘At the airport’ and ‘Travel then & now’, hardly any adjustments are needed for ‘Travel agency’, none for ‘Expectation & reality’.

sign is wished for in a certain language (e.g. French TGV with a characteristic shape).

While the picture story is suitable for all of the target languages, the signed story obviously has to be translated. For the Dicta-Sign corpus the story was originally produced in DGS and translated into written English and was then translated into each of the other target sign languages. The same holds for the video clips of the task explanations for the informants.

4. Conclusions

In the framework of the Dicta-Sign project corpus collection has so far been undertaken for DGS, LSF and GSL (BSL in preparation). Adapting the material as described made it possible to adopt it for all target languages, and a preliminary inspection of the language data collected seems to confirm our expectations of the tasks’ results. Only the transcription process now starting will allow us to analyse in detail how far our goals of “parallel” corpora have been achieved.

The length of the individual tasks as well as per session in total is roughly as it was expected, resulting in an average signing time per session (i.e. both informants, task explanations not included) for the three languages between 1:05h and 1:19h.

Feedback received from the informants so far showed that the individual tasks as well as the session as a whole were found to be interesting and appealing. For some of the tasks (esp. ‘City map’) the prerecorded task explanations were not sufficient and the moderators often needed to give further explanations.

So even at this early stage of analysis, we are convinced that, thanks to the commitment of the moderators and the motivation of the Deaf informants, we have been able to collect a corpus valuable not only for research within the project, but also to the sign language research community at large: Corpus data will be made available together with baseline transcriptions at the end of the project.

5. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231135.

6. References

- Anderson, H. et al. (1991): The HCRC Map Task Corpus. In: *Language and Speech* 34, pp. 351-366.
- Gass, S.M., Mackey, A. (2007): *Data Elicitation for Second and Foreign Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hong, S.-E., Hanke, T., König, S., Konrad, R., Langer, G., Rathmann, C. (2009): *Elicitation materials and their use in sign language linguistics*. Poster presented at the Workshop “Sign Language Corpora: Linguistic Issues” in London, July 24-25, 2009.
- Lavado, Joaquin (1991): *Guten Appetit!* Oldenburg: Lappan.

Computer-based recognition of facial expressions in ASL: From face tracking to linguistic interpretation

Nicholas Michael*, Carol Neidle†, Dimitris Metaxas*

*Computational Biomedicine Imaging & Modelling Center, Rutgers University
{nicholam, dnm}@cs.rutgers.edu

†Linguistics Program, Boston University
carol@bu.edu

Abstract

Most research in the field of sign language recognition has focused on the manual component of signing, despite the fact that there is critical grammatical information expressed through facial expressions and head gestures. We, therefore, propose a novel framework for robust tracking and analysis of nonmanual behaviors, with an application to sign language recognition. Our method uses computer vision techniques to track facial expressions and head movements from video, in order to recognize such linguistically significant expressions. The methods described here have relied crucially on the use of a linguistically annotated video corpus that is being developed, as the annotated video examples have served for training and testing our models. We apply our framework to *continuous recognition* of three classes of grammatical expressions, namely *wh*-questions, negative expressions, and topics. Our method is signer-independent, utilizing spatial pyramids and Hidden Markov Models (HMMs) to model the temporal variations of facial shape and appearance.

1. Introduction

Nowadays, speech recognition technologies have become standard components of modern operating systems, allowing average users to interact with computers verbally. Unfortunately, technology for the recognition of sign language, which is widely used by the Deaf, is not nearly as well-developed, despite its many potential benefits (Vogler and Goldenstein, 2008b; Michael et al., 2009; Neidle et al., 2009). First of all, technology that automatically translates between signed and written or spoken language would facilitate communication between signers and non-signers, thus bridging the language gap. Secondly, such technology could be used to translate sign language into computer commands, favoring the development of additional assistive technologies. Moreover, it could facilitate the efficient archiving and retrieval of video-based sign language communication and could assist with the tedious and time-consuming task of annotating sign language video data for purposes of linguistic and computer science research.

However, sign language recognition poses many challenges. First, many of the linguistic components of a sign that must be recognized occur *simultaneously* rather than sequentially. For example, one or both hands may be involved in the signing, and these may assume various hand shapes, orientations, and types of movement in different locations. At the same time, facial expression may also be involved in distinguishing signs, further complicating the recognition task. Secondly, there is variation in production of a given sign, even by a single signer. Additional variation is introduced by the *co-articulation* problem, meaning that the articulation of a sign is influenced by preceding and following signs. This can result in departures from the expected hand shape, location, and/or orientation found at the edge of a sign, and there may also be movement transitions between signs (sometimes referred to as “movement epenthesis”). Nevertheless, many methods (Vogler and Metaxas, 1998; Bauer and Kraiss, 2002; Vogler and Metaxas, 2004) have shown promising results in recogniz-

ing manual components of signs.

Furthermore, in sign language, critical grammatical information is expressed through head gestures (e.g., periodic nods and shakes) and facial expressions (e.g., raised or lowered eyebrows, eye aperture, nose wrinkles, tensing of the cheeks, and mouth expressions (Baker-Shenk, 1983; Coulter, 1979; Liddell, 1980; Neidle et al., 2000)). These linguistically significant nonmanual expressions include grammatical markings that extend over phrases to mark syntactic scope (e.g., of negation and questions). For example, in *wh*-questions (which involve phrases such as *who*, *what*, *when*, *where*, *why*, and *how*), the grammatical marking consists of lowered eyebrows and squinted eyes that occur either over the entire *wh*-question or solely over a *wh*-phrase that has moved to a sentence-final position. In addition, there may be a slight, rapid side-to-side head shake over at least part of the domain of the *wh*-question marking. With *negation*, there is a relatively slow side-to-side head shake that co-occurs with a manual sign of negation (such as NOT, NEVER), if there is one, and may extend over the scope of the negation, e.g., over the following verb phrase that is negated. The eyes may squint or close. Lastly, *topics* are characterized by raised eyebrows, wide eyes, head tilted back, and an optional nod.

Sign language recognition cannot be successful unless these nonmanual signals are also correctly detected and identified. For example, the sequence of signs JOHN BUY HOUSE could be interpreted, depending on the accompanying nonmanual markings, to mean any of the following: (i) “John bought the house.” (ii) “John did not buy the house.” (iii) “Did John buy the house?” (iv) “Did John not buy the house?” (v) “If John buys the house...”

Motivated by the importance of facial expressions and head gestures, we present a novel framework for robustly tracking and recognizing such nonmanual markings associated with *wh*-questions, *negative* sentences and *topics*. Our method extends prior work (Michael et al., 2009; Neidle et al., 2009), in which the signer’s head is tracked and appear-

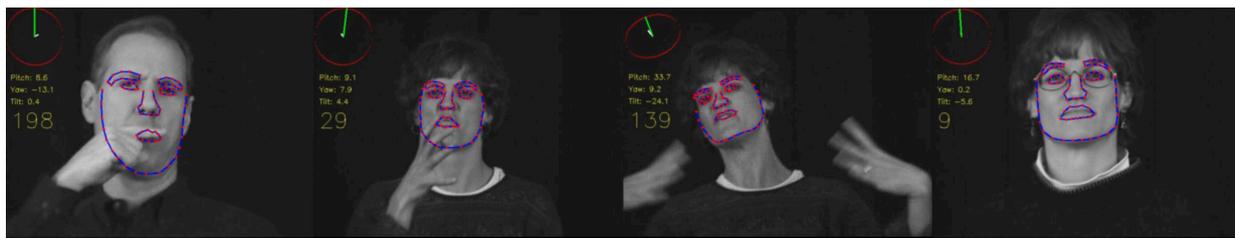


Figure 1: Sample frames showing the accuracy of tracking under challenging scenarios (partial occlusions, fast movements, and glasses), using our face tracker (Kanaujia et al., 2006). Here, red dots represent tracked landmarks. The predicted head pose is shown in the top left corner of each frame as a 3D vector

ance features, in the form of spatial pyramids (Lazebnik et al., 2006) of SIFT descriptors (Lowe, 2004), are extracted from the eye and eyebrow region (which we will refer to, henceforth, as the eye region). First, we extract additional shape features in the form of spatial pyramids of histograms of oriented gradients (PHOG) (Bosch et al., 2007). Second, we use spectral clustering (Ng et al., 2002), to reduce the dimensions of the augmented appearance and shape feature vectors. Third, by utilizing Hidden Markov Models (HMMs) (Rabiner, 1989), our method can perform *continuous* recognition in unsegmented video sequences.

2. Previous Work

As already mentioned, most research on computer-based sign language recognition has focused on the manual components of signing. A thorough review of early such efforts is presented in Pavlovic et al. (1997). Only recently have researchers begun to address the importance of facial expressions for sign recognition systems (Ong and Ranganath, 2005). An extensive review of recent developments in visual sign recognition, together with a system that captures both manual and nonmanual signs is provided by von Agris et al. (2008). However, their system requires the signer to be wearing a glove with colored markers to enable robust hand tracking and hand posture reconstruction. Additionally, in their system, the tracked facial features (lip outline, head pose, eye gaze, etc.) are not used to recognize facial expressions that have grammatical meaning. Vogler and Goldenstein (2008a; 2008b) present a 3D deformable model for face tracking, which emphasizes outlier rejection and occlusion handling at the expense of slower run time. They use their system to demonstrate the potential of face tracking for the analysis of facial expressions encountered in sign language, but they do not use it for any actual recognition. Lastly, the authors of (Michael et al., 2009; Neidle et al., 2009) use a method based on spatial pyramids (Lazebnik et al., 2006) to do *isolated* recognition of *wh-questions* and *negative* sentences. In this paper, we extend that work, so that we are now able to recognize in a *continuous* fashion *wh-questions* and *negative* sentences, as well as *topics* (i.e., no segmentation of test sequences is needed).

3. Face Tracking

Face tracking is a challenging problem because the tracker needs to generalize well to previously unseen faces and to varying illumination. It should also cope with partial occlusions and pose changes, such as head rotations, which

cause drastic changes in the shape of the face, causing it to lie on a non-linear manifold. Kanaujia et al. (2006) tackle these problems with an Active Shape Model (Cootes et al., 1995), which is a statistical model of facial shape variation, where shapes are represented by a set of facial landmarks. Through the application of Principal Component Analysis (PCA) on an aligned training set of facial shapes, they learn a model of the permissible ways in which different people’s faces differ, which is then used for face tracking.

Moreover, using a Bayesian Mixture of Experts model they are able to estimate the 3D pose of the head from the tracked landmarks. This model uses linear regressors and a multiclass classifier to map landmark configurations to predictions of head pose. Figure 1 shows the output of the ASM tracker on a few challenging input frames exhibiting rapid head movements and rotations, and partial occlusions.

Following ideas in (Michael et al., 2009; Neidle et al., 2009), the first step of our recognition framework involves tracking the signer’s head using the above described framework (Kanaujia et al., 2006), localizing the facial components (e.g., eyes, eyebrows) and predicting the 3D head pose (i.e., pitch, yaw, tilt). We then extract from the eye region the features described in the next section.

4. Feature Representation and Recognition

In order to train machine learning algorithms for recognition of facial expressions, we first need a discriminative feature representation. Therefore, we extract dense SIFT descriptors over a regular grid from the eye region of each tracked frame; these are invariant to linear transformations such as scaling and rotation (Lowe, 2004). We cluster the SIFT descriptors of a random subset of the training frames, to obtain a codebook of prototypes and then encode all other descriptors by the index of their nearest prototype.

Next, we divide each frame into imaginary grids of cells and count the relative frequency of occurrence of each encoded feature in each cell. This collection of histograms becomes the spatial pyramid SIFT representation (PSIFT). In order to measure the dissimilarity in appearance between any pair of frames, we just need to compare their PSIFT representations, essentially comparing the bins of these histograms to see how much they match, using a weighted Spatial Pyramid Match Kernel (SPMK) with the histogram intersection function (Swain and Ballard, 1991; Grauman and Darrell, 2005; Lazebnik et al., 2006).

Bosch et al. (2007) also build spatial pyramids. Instead of SIFT descriptors, their idea is to quantize the gradient

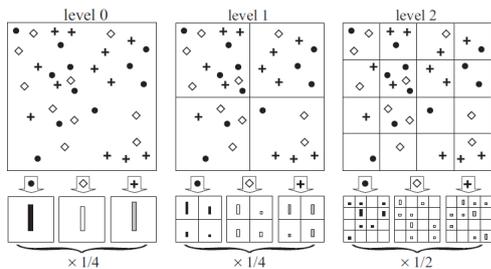


Figure 2: Toy illustration of spatial pyramid construction (Lazebnik et al., 2006), where, for simplicity, we assume there are only 3 codewords (circle, diamond, cross)

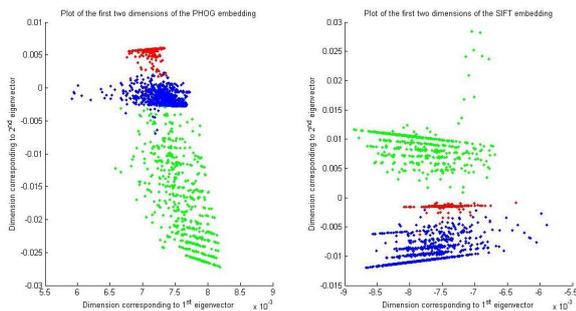


Figure 3: Spectral feature embedding of each frame (red: negative, green: topics, blue: wh-questions)

orientations of pixels into uniform bins, with each pixel’s vote being proportional to the magnitude of its gradient, forming what they call a PHOG descriptor. We compute PHOG features in the same way, but for measuring PHOG similarity we use the weighted SPMK. By combining appearance (PSIFT) and shape (PHOG) features we obtain a more discriminative representation of eye regions.

5. Recognition Models

Although combining appearance and shape features improves the discriminative power of our representation, it increases the dimensionality of our input. As such, it increases the amount of training data that we need in order to learn accurate recognition models, and this also causes an increase in complexity, thus slowing down computations.

Spectral clustering (Ng et al., 2002) is a popular method of dimensionality reduction. The feature vector of each training example is represented as a node in a graph that is connected with a weighted edge to its nearest neighbors in the training set (weights reflect degree of similarity). The algorithm then applies an eigenvalue decomposition on the matrix representing this graph, reducing the feature vector dimensionality in a way that preserves the neighborhood structure. We use SPMK as the similarity measure and reduce the dimension of PSIFT and PHOG features separately. Figure 3 shows the resulting embedding of the training set, where we see that the classes are well separated.

The final feature descriptors per frame are the combined SIFT and HOG features of reduced dimensionality together with the 3D head pose and its first order derivatives. These are used to train HMM models (Rabiner, 1989). An HMM

	None	Negative	Topic	Wh-Q
Training	10144	997	1604	1208
Testing	9359	1053	1248	1182

Table 1: Dataset composition (number of frames per class)

	Predicted Class			
	None	Negative	Topic	Wh-Q
True None	92.8%	2.9%	2.2%	2.1%
True Negative	7.7%	80.3%	5.8%	6.2%
True Topic	9.2%	4.5%	81.2%	5.1%
True Wh-Q	8.3%	5.3%	4.5%	81.9%

Table 2: Confusion matrix of HMM continuous recognition

is a probabilistic model popular for time series data, consisting of a set of hidden states. At each time step, it transitions state based on a transition probability and it emits an observation. For our recognition task, we divide frames from each training sequence into four sets, one for each class of expressions we want to recognize. We train a separate HMM for each class, using sequences segmented by class.

6. Use of the Annotated Video Corpus

The machine learning fundamental to our approach has been carried out using a linguistically annotated corpus of ASL (as produced by native signers) created at Boston University. This publicly available corpus, including 15 short narratives plus hundreds of additional elicited utterances, includes multiple synchronized views of the signing (generally 2 stereoscopic front views plus a side view and a close-up of the face), which have been linguistically annotated using SignStream™ (Neidle, 2002; Neidle et al., 2001) software, which enables identification of the start and end points of the manual and nonmanual components of the signing. Annotation conventions are documented (Neidle, 2002/2007), and the annotations are available in XML.

In order for pattern recognition algorithms to correctly identify a class of interest, they must be trained with both positive examples and negative examples. These are easily obtainable from the annotated corpus. From this corpus we selected a training set of 77 video clips of isolated utterances (negative: 17, topic: 40, wh: 20). Our testing set contained 70 such clips (negative: 15, topic: 38, wh: 17). The exact composition of these sets, in terms of numbers of frames per class, is shown in Table 1. Both sets contained three different signers. Using the methods described in previous sections, we tracked the signer’s head, extracting pose, PHOG and PSIFT features, the dimensionality of which was then reduced using spectral clustering. We then trained class-specific HMMs, optimized to recognize frame sequences of their class. To evaluate their performance at continuous recognition, we used a sliding window approach. We fed subsequences of all *unsegmented* test sequences to each HMM, classifying each frame as negative, topic, wh, or none, based on which HMM output had the highest probability of having generated each subsequence.

Recognition accuracy is summarized in the confusion matrix of Table 2.

7. Discussion

We presented a novel framework for robust *real time* face tracking and facial expression analysis from a single uncalibrated camera. Our feature representation comprises spatial pyramids of SIFT and HOG features, and head pose features, which are reduced in dimensionality using a spectral decomposition. We demonstrated that our framework is successful at continuous recognition of wh-questions, negative expressions, and topics in unsegmented video data.

Feature fusion will be crucial in helping to recognize classes of nonmanual markings that are only subtly different. Therefore, as part of our future research we will be looking at combining facial features and looking at intensity and temporal patterning of nonmanual gestures (in relation, as well, to manual signing).

8. Acknowledgements

We gratefully acknowledge substantial contributions to this project by Benjamin Bahan, Lana Cook, Quinn Duffy, Robert G. Lee, Dawn MacLaughlin, Joan Nash, Michael Schlang, and Norma Bowers Tourangeau at Boston University. Signstream™ programming has been carried out by David Greenfield and Iryna Zhuravlova. The research reported here has been partially funded by grants from the National Science Foundation: #CNS-04279883, #EIA-9809340, and #IIS-9912573.

9. References

- C. Baker-Shenk. 1983. A Micro-analysis of the Nonmanual Components of Questions in American Sign Language. Unpublished PhD Dissertation.
- B. Bauer and K.-F. Kraiss. 2002. Video-based sign recognition using self-organizing subunits. In *ICPR*, volume 2, pages 434–437.
- A. Bosch, A. Zisserman, and X. Munoz. 2007. Representing shape with a spatial pyramid kernel. In *CIVR*, pages 401–408.
- T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. 1995. Active shape models – their training and application. In *Comp. Vis. Image Underst.*, pages 38–59.
- G. R. Coulter. 1979. American Sign Language Typology. Unpublished PhD Dissertation.
- K. Grauman and T. Darrell. 2005. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, October.
- A. Kanaujia, Y. Huang, and D. Metaxas. 2006. Tracking facial features using mixture of point distribution models. In *ICVGIP*.
- S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178.
- S. K. Liddell. 1980. *American Sign Language Syntax*. Mouton, The Hague.
- D. G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.
- N. Michael, D. N. Metaxas, and C. Neidle. 2009. Spatial and temporal pyramids for grammatical expression recognition of American Sign Language. In *ASSETS*, pages 75–82, October.
- C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure*. MIT Press, Cambridge MA.
- C. Neidle, S. Sclaroff, and V. Athitsos. 2001. Signstream™: A tool for linguistic and computer vision research on visual-gestural language data. *Behavior Research Methods, Instruments, and Computers*, 33(3):311–320.
- C. Neidle, N. Michael, J. Nash, and D. Metaxas. 2009. A method for recognition of grammatically significant head movements and facial expressions, developed through use of a linguistically annotated video corpus. *Proc. of 21st ESSLLI Workshop on Formal Approaches to Sign Languages*, July.
- C. Neidle. 2002. Signstream™: A database tool for research on visual-gestural language. *Journal of Sign Language and Linguistics*, 4(1/2):203–214.
- C. Neidle. 2002/2007. SignStream annotation: Conventions used for the American Sign Language Linguistic Research Project. Technical report, American Sign Language Linguistic Research Project Nos. 11 and 13 (Addendum), Boston University. Also available at <http://www.bu.edu/asllrp/reports.html>.
- A. Y. Ng, M. I. Jordan, and Y. Weiss. 2002. On spectral clustering: Analysis and an algorithm. *NIPS*, 14:849–856.
- S. C. W. Ong and S. Ranganath. 2005. Automatic sign language analysis: a survey and the future beyond lexical meaning. *IEEE TPAMI*, 27(6):873–891, June.
- V. I. Pavlovic, R. Sharma, and T. S. Huang. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE TPAMI*, 19:677–695.
- L. R. Rabiner. 1989. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- M. J. Swain and D. H. Ballard. 1991. Color indexing. *IJCV*, 7:11–32.
- C. Vogler and S. Goldenstein. 2008a. Facial movement analysis in ASL. *Univers. Access Inf. Soc.*, 6(4):363–374.
- C. Vogler and S. Goldenstein. 2008b. Toward computational understanding of sign language. *Technology and Disability*, 20(2):109–119.
- C. Vogler and D. Metaxas. 1998. ASL recognition based on a coupling between HMMs and 3D motion analysis. In *ICCV*, pages 363–369.
- C. Vogler and D. Metaxas, 2004. *Handshapes and movements: Multiple-channel ASL recognition*, pages 247–258. LNAI. Springer, Berlin.
- U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. 2008. Recent developments in visual sign language recognition. *Univers. Access Inf. Soc.*, 6(4):323–362.

Organizing data in a multilingual observatory with written and signed languages.

Cédric Moreau(**), Bruno Mascret(*)

Université de Lyon *, CNRS
Université Lyon 1, LIRIS, UMR5205, F-69622, France

INS HEA **
Institut National Supérieur de Formation et de Recherche pour
l'Éducation des Jeunes Handicapés et les Enseignements Adaptés.
58, 60 av des Landes 92150 Suresnes France

UMR 7023 **
Université Paris 8
2 rue de la Liberté
93200 Saint-Denis France

cedric.moreau@inshea.fr, bruno.mascret@liris.cnrs.fr

Abstract

The Académie Française institution is assigned and devoted to defending the French language and to making it a common heritage for all French speakers. The French Sign Language (LSF) has never had such a support.

To face this situation, a reference tool has been created, supported by the French Ministry of Education and by the General Delegation to the French language and to languages of France. This tool is a collaborative website entirely bilingual French and LSF, and which proposes for each concept at least one definition and its associated descriptors in various knowledge fields. Before being spread on-line, the information given by users (text, picture, video, presentation) is examined by experts on form and content, and is validated or rejected by these experts with an explanation.

Considering regional and sociological differences, several signs may be proposed and validated for one concept. Our project does not wish to choose the "ideal sign", but wants to submit to our identified users all the proposals and to list their comments (have they come across this sign and if so, in which context). A set of information is thus collected for each sign and can be related to users profiles. The website is therefore an exchange platform, but can also be used as a linguistics observatory.

One of our main issues concerning the data organization was to manage to adjust users different viewpoints and different uses of the website. Indeed, our platform's goal is not to make a simple dictionary but to create a network of ontologies. Our other issue is now that we cannot use a rigid organization model, because our website must constantly evolve and include new concepts and new descriptions or functionalities, such as illustrations, homonyms, antonyms, etc. In this article we will first briefly describe our platform's goals, then present our specific data organization which allows for example several classifications to be used simultaneously. We will illustrate this approach's interest with a critique of Dewey's classification, that we had at first implemented despite its limits (acceding to a precise concept is difficult, the organization is not intuitive, recent concepts or specific LSF concepts cannot be referenced, etc.). We will propose to replace it with classifications directly created by our users and corresponding to their expectations and needs. This way the tree diagram is built gradually and supervised by experts in each knowledge field.

Each content thus goes with descriptors and classifiers allowing it to play different parts depending on the context. Therefore a content can at the same time be a concept, a classification theme or sub-theme, or an illustration – the context will mobilize the appropriate contents depending on their descriptors and classifiers.

We will finally present our current work on integrating direct resources in LSF through descriptors defining a sign's spatial position and its moves (hands, body and face), in order to highlight our platform's great ability to evolve. We will also show that this data organization allows an easy conversion to other countries' sign languages.

Key words: French Sign Language, LSF, written languages, dynamic classification, deaf, collaborative website, concept, ontology.

1. Preamble

According to Gillots' official report¹, 80% of the French deaf people are illiterate, and only 5% reach higher education. Dalle has also declared in (Dalle, 2003) that "illiteracy, short knowledge of written French, lack of diploma and of qualifications as well as communicating problems have great consequences on deaf adults' social and professional integration".

Since the French 11th February 2005 act, public schools cannot refuse for any reason to take in a child living in its defined area. Besides, all public websites must be entirely accessible. In 2010, a state diploma will be created for LSF teachers. In this context, new needs have appeared, and appropriate bilingual teaching tools are increasingly demanded.

Just as the French language, LSF has many regional vocabulary differences, and it constantly enriches itself with new words, thanks to its speakers who create signs to name new

¹<http://cis.gouv.fr/spip.php?article1516>

concepts and/or concepts that are specific to a knowledge field. As the deaf people increasingly access university and professional environments, this phenomenon is enhanced. (Duquesne-Belfais, 2007) stresses out that each concept – once its has been attributed a name or a sign – can be used to define a more abstract concepts characteristics and can take part in building a knowledge network. Nominalizing the concepts characteristics allows it to change its status, switching from implicit to explicit, and to take part in constructing a rigorous language.

2. "OCELLES" PROJECT

The main support of the "OCELLES" project (Moreau, 2008) is a collaborative website, entirely bilingual French/LSF, and which proposes for each concept at least one definition and its associated descriptors – in both languages – in all possible knowledge fields. Before being spread on-line, the information given by users (text, picture, video, presentation) is examined by experts on form and content, and is validated or rejected by these experts with an explanation.

The project is currently under testing and will be published at <http://www.ocelles.fr>.

2.1. Managing users and rights

Running with a GPL licence², the website is free to all. Users may access different statuses:

- visitors browse on the websites public content.
- they become users when registering and filling a form establishing their profile. The collected information – on their scholarship, track record, languages used within family and social lives – will balance their answers and advice concerning proposed signs.
- writers are users who have accepted the publication terms. They may propose new contents and concepts to experts – possibly supported by a classification. They may also add videos to illustrate other authors text sequences.
- experts are writers who validate the contents deposited by writers. They must also provide an explanation to the writer in the event of non validation.
- the administrator is expert in all fields. However his main role is managing the portal without taking part in expertises and publications.

We must specify that writer and expert statuses are only attributed to a users knowledge field: a user may be expert in mathematics, writer in philosophy and plain user in all other fields.

Managing users rights so precisely should lead to a democratic and community use of our website, because the hierarchy only depends on the chosen theme.

2.2. Concepts

For each concept, a specific and dynamic webpage proposes a definition, a translation into written language and one or several signs in Sign Language. Considering regional and sociological differences, several signs may be proposed for each concept. Our project does not wish to choose the "ideal sign", but wants to submit all proposals to users. The definitions respect the following rules: Written definitions must:

- give the concepts meaning and its main characteristics,
- be precise,
- begin with a general explanation,
- be a suitable substitute to the unknown word in a text,
- not integrate other definitions,
- not contains other words having the same root,
- not be circular.

Illustrations, examples, comments (educational or linguistic etc.), slight differences, regional uses, connotations and other are in addition to the definition, not a part of it. A Sign Language definition:

- must not contain any regional code or name (however local signs may be used to refer to a concept),
- must avoid neologisms,
- should be punctuated in order to enable its sequencing.

The signing flow is adjusted for deaf or hearing learners.



Figure 1: Screenshot of a "concept page".

The platforms goal is not to provide a plain dictionary, but a real network of ontologies. Links enable associations between concepts – ex. "thesis" refers to "arguing" – thanks to the "see also", "close concepts", and "opposite concepts" descriptors. The links between concepts are flexible, and more type of links can easily be integrated to "OCELLES".

²<http://www.gnu.org/licenses/licenses.html>

2.3. signs

Each sign proposal opens a new web page. Examples, context and other descriptors as well as linguistic and epistemological comments can be added. Users are encouraged to answer questions about:

- the context(s) in which they have encountered each sign (class, job etc.),
- the sign characteristics – i.e. formed with one or two hands, coming from a transfer (Cuxac, 2000) in form, situation, person, configuration, position, moves, facial expression etc.

2.4. Proposals summary

For each concept, the answers given by other users are gathered and summed up on one page. They can be linked to their profiles, thus enabling for example a collection of information about geographical localization of each sign. More than an exchange platform, the website is also a synchronic linguistics observatory.

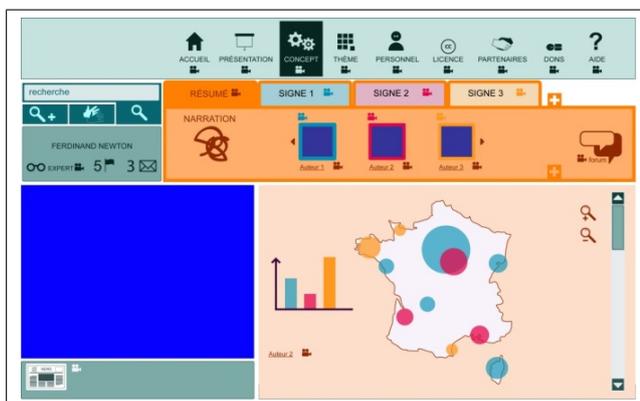


Figure 2: Screenshot of a "summary page".

3. Data organization model

3.1. A rigid classification and its limitations

At first we have used Dewey's decimal classification (DDC), usually used in libraries. We have chosen this system, developed by Melvil Dewey in 1876³, because it exists and classifies the whole set of human knowledge.

However, user tests, made on both deaf and hearing persons, rapidly pointed up the difficulties we had sensed. The DDC consists in classifying works and knowledge into 10 general categories, each one of them being divided and subdivided each time into 10 subcategories and so on as many times as needed. Looking for a specific concept through this tree diagram makes it imperative to:

- know into which categories and subcategories the concept will be classified,
- make no mistake through the tree diagram in choosing the subdivisions.

³<http://www.oclc.org/dewey/>

This approach, not intuitive, will locate a low-level concept very far from the head of the diagram, thus the low-level concepts will be found only if the high-level ones are known and understood.

As for the linguistic system LSF, it combines a categorical aspect with its vocabulary structure. According to (Courtin, 1998), the use of LSF by deaf children whose parents are deaf and already signers increases categorizing abilities compared to hearing or oralizing children. Courtin has observed this phenomenon especially when the categorization respects our world's complexity by using prototypes or diagrams (Bideaud et al., 1993). Indeed, signers often refer to a concept through a series of prototypical examples of it, thus defining the concept by extension. A rigid and arbitrary classification could then disturb deaf users.

Besides, where and how should new concepts directly stemming from Sign Language be classified, in a rigid classification set upon written language? (ex. "LS Video", video recordings of formalised LSF used as a differed communication, or "signary", set of all signs in Sign Language).

3.2. Dynamic classification

One of our main issues concerning the data organization is to manage to adjust users' different viewpoints and different uses of the website. Our portal must be able to easily evolve and include new concepts as well as new descriptions or functionalities, such as illustrations, homonyms, antonyms, etc. Keeping using a rigid data organization is impossible. That is why we have chosen a data organization which considers a priori each one of the website's elements as a content. In parallel, an associative and dynamic structure has been set up, enabling to link contents together according to their roles and to the descriptors associated to these roles (Moreau and Mascret, 2008).

This way, one content may be used several times because in different contexts, depending on the associations it belongs to – roles and descriptors (Bénel, 2003).

Let us give an example: in our website, a classification node has a role of theme. A theme is also a concept for example "language". This theme lists other themes and concepts. "Language", as a theme, contains the themes "lexicon", "grammar"... Moreover, "language", as a content, also has a role of illustrator to the concept "lexicon". This way, one content – here "language" – has different roles (theme, concept). Each one of these roles has its own descriptors (concept's illustration, other related concept...).

Finally, the diagram tree must allow a concept to be classified in several themes without duplicating it. Libraries often face this problem when classifying works containing several themes – a book about science in the 19th century should be classified into history as well as science.

3.3. Discussion: a dynamic classification built on LSF linguistic parameters?

According to (Cuxac, 2000), two discursive enunciation strategies coexist in LSF. Through the visual-gestural channel, the signer chooses to say without showing, or to say and show. This way, he can visually re-present the experience thanks to the greatest resemblance between a sequence of signs and the experience itself. Or else he can use the

standard sign that does not resemble the referent. Based on this theory of iconicity, our research draws the assumption of a hierarchy between the linguistic parameters used in signs as meaningful elements.

If those greatly iconic structures involve infra-lexical linguistic elements that do not belong to the lexicon, they appear most often in narrating sequences and remain nonetheless unmentioned in Sign Language dictionaries. However, if we consider that these structures are an integral part of Sign Language, how should we integrate them into our web site?

Two perspectives are suggested to answer this question. The first one consists in considering the minimal structures of realization in Sign Language. The linguistic parameters of configuration, movement, location (Stokoe et al., 2000) and orientation (Friedman, 1977), (Liddell, 1980), (Moody, 1983), (Yau, 1992) cannot be considered as such. Indeed, even if a human mind can make a relevant distinction between them as isolated elements conveying meaning, they must be used simultaneously in order to be activated while communicating. Contrary to vocal languages, realizing a signifying form in a Sign Language cannot be made through a succession of distinct realizations of isolated and non-signifying elements. Minimal realization structures in Sign Language may be ranged on a growing complexity scale, starting from the formal transfer (infra-conceptual level) and going up to the double transfer (level where several actors, location parameters and utterances can be combined). These various structures use the same linguistic parameters during the same realization laps of time.

The second perspective is based on the dialectics between syntagm and paradigm. When narratives contain highly iconic structures, the value of an element at a given time undergoes a type of syntagmatic pressure – which does not necessarily come from preceding or following units, but from other units occurring at the same time and taking part in the minimal form of realization as well. Yet, the simultaneity is not a sufficient clue to conclude that it is a paradigm, since this pressure can be seen. In a Sign Language, the pressure stemming from the context does not only influence the temporal dimension. The spatial dimension exerts constraints as well, but this time simultaneous instead of successive. Regarding these two perspectives, our users are questioned about their perceptions and representations of the meaningful infra-conceptual units – while first visiting each “sign page”.

We do not want to collect “correct” answers, but to gather the most identical ones. This way, our classification leans on a consensus amongst users. However, our experts can impose a classification and may concentrate the researches for a sign through these answers, without necessarily using the material as a final classification.

Based on our users’ answers, descriptors and/or classifiers are assigned to each sign, according to the summary of a dynamic amount of identical and meaningful answers. The data base model we propose is based on the idea of modelling the interactions giving sense to the content – and not the content itself. The polymorphic use of contents implies a data organization based on the role we wish a content to play, as explained above. In this way, an “answer” – as

a content – has both roles of answer to a question and of classifying and research element. One or several specific descriptors correspond to each role.

This approach of a dynamic classification of concepts, built upon LSF linguistic parameters specific to each sign, enables us to propose our users to look up concepts through the site directly in LSF, without having to know the concepts’ written signifiers. Later on, a dynamic classification could also be based upon sign writing⁴.

4. References

- A. Bénel. 2003. *Consultation assistée par ordinateur de la documentation en Sciences Humaines : Considérations épistémologiques, solutions opératoires et applications l’archéologie*. Ph.D. thesis. Thèse de doctorat en informatique, Institut National des Sciences Appliquées de Lyon, 2003.
- J. Bideaud, O. Houde, and J.-L. Pedinielli. 1993. *L’homme en développement*. PUF, Paris.
- C. Courtin. 1998. *Surdit , langue des signes et d veloppement cognitif*. Ph.D. thesis, Universit  Paris V.
- C. Cuxac. 2000. *La Langue des Signes Francaise : les voies de l’iconicit *. OPHRY, PARIS.
- P. Dalle. 2003. La place de langue des signes dans le milieu institutionnel de l ducation: enjeux, blocage et  volution. *Langue franaise*, 137.
- F. Duquesne-Belfais. 2007. *Activit  et langages dans la conceptualisation math matique : des apprentissages des  l ves sourds la formation de leurs enseignants*. Ph.D. thesis, Universit  Lille 1.
- Lynn A. Friedman. 1977. *On the other hand : new perspectives on American sign language*. Academic Press, New York.
- S.K. Liddell. 1980. *American Sign Language Syntax*. The Hague.
- B. Moody. 1983. *Histoire et Grammaire, tome I*. Ellipses, Paris.
- C. Moreau and B. Mascret. 2008. Lexique LSF. In Onno et al (eds) Crasborn, editor, *3rd Workshop on the Representation and Processing of Sign Languages. 6th International Conference on Language Resources and Evaluation, LREC 2008*, Construction and Exploitation of Sign Language Corpora., pages 138–140, June.
- C. Moreau. 2008. Lexiquelsf : vers une web acad mie de la langue des signes franaise. *La nouvelle revue de l’adaptation et de la scolarisation*, 43, October.
- W.C. Stokoe, D.C. Casterline, and C.G. Croneberg. 2000. In *A dictionary of american sign language*, WASHINGTON D.C. Gallaudet College Press.
- S.-C. Yau. 1992. *Cr ation Gestuelle et d but du Langage Cr ation de langues gestuelles chez les sourds isol s*.  ditions Langues Crois s, Hong Kong.

⁴<http://www.signwriting.org/>

Building a Sign Language corpus for use in Machine Translation

Sara Morrissey, Harold Somers, Robert Smith, Shane Gilchrist and Sandipan Dandapat

Centre for Next Generation Localisation

Dublin City University

Glasnevin, Dublin 9, Ireland

{smorri,hsomers,rsmith,sdandapat}@computing.dcu.ie,shane.gilchrist@gmail.com

Abstract

In recent years data-driven methods of machine translation (MT) have overtaken rule-based approaches as the predominant means of automatically translating between languages. A pre-requisite for such an approach is a parallel corpus of the source and target languages. Technological developments in sign language (SL) capturing, analysis and processing tools now mean that SL corpora are becoming increasingly available. With transcription and language analysis tools being mainly designed and used for linguistic purposes, we describe the process of creating a multimedia parallel corpus specifically for the purposes of English to Irish Sign Language (ISL) MT. As part of our larger project on localisation, our research is focussed on developing assistive technology for patients with limited English in the domain of healthcare. Focussing on the first point of contact a patient has with a GP's office, the medical secretary, we sought to develop a corpus from the dialogue between the two parties when scheduling an appointment. Throughout the development process we have created one parallel corpus in six different modalities from this initial dialogue. In this paper we discuss the multi-stage process of the development of this parallel corpus as individual and interdependent entities, both for our own MT purposes and their usefulness in the wider MT and SL research domains.

1 Introduction

This paper describes the planning and construction of a multimedia parallel corpus for the purpose of developing a machine translation (MT)-based approach to using technology to assist patients with limited English in a healthcare scenario. Focussing on the first point of contact a patient has with a GP's office, the medical secretary (receptionist), we are developing a corpus representing the dialogue between the two parties when scheduling an appointment. The corpus is a multimedia six-way parallel corpus consisting of (a) audio recordings of the original material, (b) written English transcription, translated into (c) Irish Sign Language (ISL) video recordings and (d) Bangla text. From the video recordings, transcriptions in (e) HamNoSys and (f) the corresponding SiGML notations have been made, the last of these being suitable to generate ISL with an animated computer figure (avatar). Each of these elements is discussed in this paper.

1.1 Assistive technology and appointment scheduling

There is no shortage of literature confirming that lack of knowledge of the host country's language and the ensuing communication difficulties constitute the single most important barrier to healthcare (e.g Jones & Gill, 1998; and many others), and an equally rich literature, which we will not review here, discusses traditional ways of addressing this problem, through use of interpreters and other services. While this observation usually applies to refugees and other immigrants, it applies equally to Deaf people (e.g. McEwen and Anton-Culver, 1988; and many others). On-going research has been investigating the use of various types of language technology to address this problem for oral languages, including (but not restricted to) MT (Somers and Lovel, 2004; Somers, 2006). In the field of spoken-language MT, cooperative goal-oriented dialogues such as appointment scheduling have always been the most widely targeted dialogue type, while the medical domain has become an important focus of research for speech translation, with its own specialist

conferences (e.g. at HLT/NAACL06 in New York, and at Coling 2008 in Manchester).

1.2 SL translation

SL MT is in the early stages of development, in comparison with mainstream MT. Widespread documented research in SL MT did not emerge until the early 1990s. This is understandable given the comparatively late linguistic analysis of SLs (Stokoe, 1960). Despite this, and within the short time-frame of research, the development of systems has roughly followed that of spoken language MT from rule-based approaches toward data-driven approaches.

Rule-based systems, such as the Zardoz system (Veale et al., 1998) and the ViSiCAST project (Marshall and Sáfár, 2002, 2003) carry out a deep linguistic analysis on a syntactic and sometimes semantic level in order to define rules for translation. More recent systems developed at RWTH Aachen University (Dreuw et al., 2007) and Dublin City University (Morrissey, 2008) have employed data-driven approaches that eschew heavy linguistic analysis in favour of empirical and statistical data. Both methodologies are heavily dependent on the suitability of the transcription approach chosen.

In the remainder of this paper we discuss our methods and the issues and problems in each stage of the corpus building activity, ending with a preview of our intended uses of the corpus.

2 Elicitation method

Our first task was to collect an English-language corpus of patient-receptionist dialogues. A major difficulty in gathering genuine data in the medical field, or any domain where personal information is involved, is that the confidentiality and other ethical issues more or less preclude using genuine data collected *in situ*. This difficulty has long been recognised in medical training, where "standardized patients" (SPs) are used with medical students, that is, actors trained to simulate

consistently the responses of a patient in a particular medical setting. Barrows (1993) describes some of the pros and cons of using SPs. As far as we could ascertain, no reported study has used SPs only for appointment scheduling, though this activity has been a (usually minor) part of many studies. Training SPs is of course a major undertaking in itself necessarily involving experienced experts, so for the purposes of this project we made a compromise in that we engaged an experienced GP's receptionist to participate in a number of role-play sessions with the native English speakers among the authors (HS, SM, RS). These were all recorded and later transcribed. Following the receptionist's guidance, we role-played a number of scenarios:

- general appointment scheduling with the GP or practice nurse, including scheduling on behalf of a third party (a child, an old person, or someone who doesn't speak English),
- emergency situations
- scheduling of specific activities, e.g. vaccinations, bringing in samples, collecting results, having stitches removed, etc.
- changing or cancelling appointments

Many of the dialogues involved negotiations of a general nature (e.g. exploring available days and times) or more specific to the individual person or purpose. In each case, the receptionist made suggestions based on her real-life experience of types of interactions that had not already been covered. In this way, we believe that our corpus contains samples that are realistic, and offer a broad coverage of our target domain, even if they are not genuine in the literal sense.

Our recordings comprise 350 dialogue turns. In transcription, this works out at just under 3,000 words (a very small corpus by any standards), each dialogue turn on average roughly 8 words.

3 Translation

The next stage in the process was to translate our English corpus into ISL (and Bangla). ISL is the main SL used in Ireland's Deaf community. Historically, Deaf children were taught separately according to their sex, leading to the rise of two main variants in ISL, i.e. male signs and female signs. Among the younger generation, there has been an acceleration in contact between varieties due to increasing social interactions between males and females, and thus contemporary ISL could be said to include both dialects. Older members of the community may not be familiar with variants from the other side.

Signed English (SE), promoted by a Deaf school in Dublin, is used by a number of Deaf people in the greater Dublin area, especially among the older generation. It is seen by some as prestigious, despite the more recent view that ISL is the way forward. There is a strong link between SE and the Church: for example the Lord's Prayer and Hail Mary are done in SE rather than ISL.

For the present project, a Deaf consultant was engaged to discuss the most suitable strategy. It was agreed that Deaf people who use SE are capable of following ISL no matter how fluent it is. On the other hand, native signers of ISL would have trouble following SE. It can be argued that SE is part of ISL (just as finger spelling is). In this context, when discussing ISL, we are talking about a register where there is very little influence from English and this

in turn provides a challenge for translating since ISL is a minority language used in face-to-face communication while English is used when writing and reading. However, low levels of English literacy among Deaf people is a major motivation for this project, so it was agreed that our translations into ISL should show a minimal influence from English.

3.1 Challenges in translation

Translation between any languages, whether related or not, involves cases where closely following the source text (a "literal" translation, within the grammatical constraints of the target language) can result in a stilted, unnatural or, in the worst case, unacceptable translation. This is especially the case when translating between English and ISL which differ both typologically and (obviously) in the medium of expression.

A particular difference is the role of pragmatics in the two languages. ISL utterances tend to reflect the immediate context much more explicitly than English, so that it is difficult to provide an ISL translation of a given dialogue turn out of context. This also has serious implications for our approach to MT.

A good example is the dialogue in (1):

- (1) A. Which doctor would you prefer?
B. I don't mind.

In ISL, A will depend on how many choices there are: if there are three people, they will first have to be identified, using the neutral space to show three different placements. Then <WHICH?> is signed,¹ spreading it across the neutral space. For the response B, the signer would just point at each placement then sign <EITHER>, then <DON'T MIND>. But just signing <DON'T MIND> without the context would be misleading or meaningless.

Interestingly, this exchange posed a similar problem for translation into Bangla where a literal translation (2a) is less preferable than a more explicit translation (2b).

- (2) a. আমি কিছু মনে করব না।
āmi kichhu mane karaba nā
I don't mind.
b. যে কোনো একজনকে দেখলেই হবে।
ye kono ekajanake dekhālei habe
Can see either of them.

Open-ended questions in English are better translated into ISL with a range of possible answers. For example, we translated (3a) as (3b).

- (3) a. How long will it take you to get here?
b. YOU-GET-HERE WHAT TIME? 10 MINUTES? 5 MINUTES?.

The strategy of "explicitation" is well known in translation studies (Klaudy, 1998). There are many examples of this in our corpus: for many conditions the sign includes location on the body, for example <PAIN> or <RASH>, the sign for which should indicate whether the condition is on the arm, on the back, on the face etc. One tactic, though against our principle of providing natural translations, is to fingerspell <R-A-S-H>.

4 Video recording

Although a number of SL video corpora have been collected, there are no agreed standard formats, often

¹ Our convention in this paper is to indicate signs with an approximate English gloss in small capitals.

because of differences in the underlying purpose behind the corpora.

The first batch of signing was recorded using an analogue TV camera at the DCU TV studio using miniDV tapes. Upon advice from technical staff at DCU School of Communication, for the remainder a Sony XCAM HDD digital camera was used. This resulted in a big jump in quality and ease of editing. The first batch was transferred to file using the DV deck which was highly time consuming and the quality was not good. The second batch showed a vast improvement in comparison.

Following the lead of the Signs of Ireland corpus project (Leeson and Nolan, 2008), the individual recordings were stored as .MOV files. They were edited using the Final Cut Pro video editing program on a Apple iMac G5 at the DCU School of Communication

Three days were spent translating the English sentences into ISL: often some trial and error was needed to arrive at a translation that was satisfactory.

After the initial recording session, our Deaf consultant reviewed the translations. Approximately 90 of the 350 sentences had to be redone for several reasons because they were felt to be too close to the English, because facial expressions were not appropriate, placement and neutral space not used correctly, and other performance frailties due to the signer's fatigue towards the end.

In retrospect, it probably would have saved effort if the reviewer had been present during the original recordings. Despite the budgetary implications, this would have saved time and energy, and would have improved the overall quality of the corpus.

This highlights one of the most interesting differences between translation into SLs and oral languages: because of the "performance" element of the SL, the step equivalent to revision in the (oral language) translation flow is considerably more demanding.

5 Transcription

The next stage was to transcribe the videos into a form suitable for textual manipulation. It is probably not necessary in the present forum to justify our use of a transcription that reflects the actual signs in a more explicit way than the widely used convention of glossing into quasi-English, even if that representation method is advantageous for ready reference, as in our discussion in the previous section.

Our choice here was guided by our main purpose, ultimately, to use the corpus of translations in a data-driven MT system to generate translations of (novel) English inputs as simulations of ISL using a computer graphic animated character (avatar).

After looking at several alternatives, it was decided to use the Hamburg Notation System (HamNoSys) and its related mark-up language SiGML.

5.1 HamNoSys

HamNoSys is a well-established transcription system developed by the Institute for German Sign Language and Deaf Communication at the University of Hamburg for all SLs (Prillwitz et al., 1989). HamNoSys is a phonetic notation system purpose-built for use by linguists in their detailed analytical representation of signs and sign phrases rather than as a writing system for SLs. According to Bentele (n.d.), it consists of about 200 symbols

covering the parameters of hand shape, hand configuration, location and movement. The symbols are iconic so as to be more easily recognizable and learnable. The order of the symbols within a string is somewhat fixed, but it is still possible to transcribe a given sign in lots of different ways. The notation is essentially phonemic, so the transcriptions are very precise, but on the other hand also very long and cumbersome to decipher. Without doubt, the learning curve for a newcomer to HamNoSys is relatively steep.

Transcribing HamNoSys is all the more arduous because the most widely used annotation tool, ELAN,² does not handle HamNoSys. To our knowledge, the only transcription software available for HamNoSys that allows alignment with the video timestamp is iLex (Hanke, 2002), though we have not yet got access to this tool.

5.2 SiGML

Closely associated with HamNoSys is SiGML (Signing Gesture Mark-up Language) (Elliott et al., 2004), a form of XML which defines a set of XML tags for each phonetic symbol in HamNoSys. SiGML files are represented as plain text which means they can be easily handled by computer, e.g. for transmission, and by the MT system (see below). SiGML was developed by the Virtual Humans group at the University of East Anglia over a three year period to support the work of the EU-funded projects ViSiCAST (Elliott et al. 2000; Kennaway, 2001, 2003) and eSIGN (Kennaway et al., 2007), whose main focus was to provide communication tools in the form of computer-graphic animated figures (avatars) for members of the Deaf community.

The SiGML representation of the HamNoSys notation of the SL sequence is readable by the AnimGen 3D rendering software (Kennaway, 2003).

6 Avatars

Research into synthesising SLs is still in the early stages of development. Most existing systems use avatars to synthesise sign language in real-time (e.g. Grieve-Smith, 1999; Krňoul et al. 2007). Using a tool called eSIGNeditor (Kennaway et al., 2007) developed during the eSIGN project, we are able to compose HamNoSys scripts for the corpus and validate them in real-time by using the processing pipeline for synthetic SL generation also developed in the eSigns project. Using this system, it is not possible however to align the HamNoSys transcriptions to the time stamps on the video files as it would be with iLex.

State-of-the-art SL synthesis can be compared to the somewhat robotic and artificial nature of early speech synthesis output. Current problems with the avatar include the need for better collision detection, more naturalness and less jerkiness. Collision detection is a means to incorporate awareness of the physical space taken up by the human body. Getting the avatar to position its hands exactly where you want them, for example close to the face, requires quite subtle programming: by default the hands and arms will take the shortest route possible to their destination, sometimes passing through another part of the body. There is a trade-off between collision

² <http://www.lat-mpi.eu/tools/elan/>, accessed 20.3.10

isation workflows (Du et al., 2010). Either the HamNoSys transcription or the SiGML code could form the text-based version of ISL required for MT processing. Both provide a level of granularity much finer than the usual approach to EBMT, which is usually based mainly on word-based matches, rarely on letter strings. It will be interesting, and a matter of research, to see the effect this has on the alignment and recombination phases of EBMT. For example subtle differences between signs that give different nuances of meaning and expression, for example in hand position, movement, or shape, will be captured by the system and used in the translation.

Using SiGML allows us to maintain the phonetic description of the signs required for animation by the avatar and avoids the use of glossing and other techniques that can misrepresent the language.

While current research efforts are focussed on English-to-ISL MT, we hope to expand the system in the future to include recognition components to allow for ISL-to-English MT, and thus a complete bidirectional translation system.

Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (<http://www.cngl.ie>) at Dublin City University. We would also like to acknowledge the contributions to this work of Eoin Campbell and Damien Hickey of the DCU School of Communication, Suzanne Lindfield, and Alvean Jones, and the assistance of our colleagues at UEA, Norwich.

References

- Barrows, H.S. (1993) An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Academic Medicine* 68, pp. 443–451.
- Bentele, S. (n.d.) About the HamNoSys system. <http://www.signwriting.org/forums/linguistics/ling007.html>, accessed 20.3.10.
- Dreuw, P., D. Stein and H. Ney. (2007) Enhancing a Sign Language translation system with vision-based features. In M. Sales Dias, S. Gibet, M.M. Wanderley and R. Bastos (eds) *Gesture-Based Human-Computer Interaction and Simulation, 7th International Gesture Workshop, GW 2007, Lisbon, Portugal, Revised Selected Papers* (LNAI 5085), Berlin (2009): Springer, pp. 108–113.
- Du, J., He, Y., Penkale, S. and Way, A. (2009) MaTrEx: the DCU MT System for WMT 2009. In *EACL 2009 Fourth Workshop on Statistical Machine Translation*, Athens, pp 95–99.
- Du, J., Roturier, J. and Way, A. (2010) TMX markup: A challenge when adapting SMT to a localisation environment. Paper submitted to 14th Annual Conference of the European Association for Machine Translation, Saint-Raphaël, France.
- Elliott, R., Glauert, J.R.W., Jennings, V. and Kennaway, J.R. (2004) An overview of the SiGML notation and SiGMLSigning software system. In *Fourth International Conference on Language Resources and Evaluation, LREC 2004*, Lisbon, pp. 98–104.
- Elliott, R., Glauert, J.R.W., Kennaway, J.R. and Marshall, I. (2000) The development of language processing support for the ViSiCAST project. In *ACM SIGACCESS Conference on Computers and Accessibility, Proceedings of the fourth international ACM conference on assistive technologies*, Arlington, Virginia, pp. 101–108.
- Hanke, T. (2002) iLex – A tool for sign language lexicography and corpus analysis. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, pp. 923–926
- Jones, D. and Gill, P. (1998). Breaking down language barriers. *British Medical Journal* 316 (7127), pp. 1476–1480.
- Kennaway, R. (2001) Synthetic animation of deaf signing gestures. In I. Wachsmuth and T. Sowa (eds) *Gesture and Sign Language in Human-Computer Interaction, International Gesture Workshop, GW 2001, London, UK, 2001, Revised papers*, (LNCS 2298), Berlin (2002): Springer, pp.149–174.
- Kennaway, R. (2003) Experience with and requirements for a gesture description language for synthetic animation. In A. Camurri and G. Volpe (eds) *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop, GW 2003, Genova, Italy, 2003, Selected revised papers*, (LNAI 2915), Berlin (2004): Springer, pp. 300–311.
- Kennaway, J.R., Glauert, J.R.W. and Zwitterlood, I. (2007) Providing signed content on the Internet by synthesized animation. *ACM Transactions on Computer-Human Interaction* 14, pp. 1–29.
- Klaudy, K. (1998) Explication. In M. Baker (ed.) *Encyclopedia of Translation Studies*, London: Routledge, pp. 80–85.
- Krňoul, Z., Kanis, J., Želený, M. and Müller, L. (2007) Czech text-to-sign speech synthesizer. In A. Popescu-Belis, S. Renals and H. Bourlard (eds) *Machine Learning for Multimodal Interaction, 4th International Workshop, MLMI 2007, Brno, Czech Republic, 2007, Revised selected papers* (LNCS 4892), Berlin (2008): Springer, pp. 180–191.
- Leeson, L. and Nolan, B. (2008) Digital Deployment of the Signs of Ireland Corpus in Elearning. In *3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora*, Marrakech, Morocco, pp. 112–121.
- Ma, Y., Okita, T., Çetinoğlu, Ö, Du, J. and Way, A. (2009) Low-resource Machine Translation using MaTrEx: the DCU Machine Translation System for IWSLT 2009. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2009)*, Tokyo, pp.29–36.
- Marshall, I. and Sáfár, E. (2002) Sign language generation using HPSG. In *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-02)*, Keihanna, Japan, pp. 105–114.
- Marshall, I. and Sáfár, E. (2003) A prototype text to British Sign Language (BSL) translation system. In *41st Annual Meeting of the Association of Computational Linguistics*, Sapporo, Japan, pp. 113–116.
- McEwen, E. and Anton-Culver, H. (1988) The medical communication of deaf patients. *Journal of Family Practice* 26, pp. 289–291.

- Morrissey, S. (2008). Data-driven machine translation for sign languages. PhD Thesis, Dublin City University, Dublin.
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T. and Henning, J. (1989) *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide*. Hamburg: Signum.
- Somers, H. (2006) Language engineering and the pathway to healthcare: A user-oriented view. In *HLT/NAACL-06 Medical Speech Translation, Proceedings of the Workshop*, New York, NY, pp. 32–39.
- Somers, H., Dandapat, S. and Naskar, S. (2009) A review of EBMT using proportional analogies. In *Proceedings of 3rd Workshop on Example-Based Machine Translation*, Dublin, pp. 53–60.
- Somers, H. and Fernandez Diaz, G. (2004) Translation Memory vs. Example-based MT: What is the difference? *International Journal of Translation* 16 (2), pp. 5–33.
- Somers, H. and Lovel, H. (2003) Computer-based support for patients with limited English. In *Association for Computational Linguistics EACL 2003, 10th Conference of the European Chapter, Proceedings of the 7th International EAMT Workshop on MT and other language technology tools: Improving MT through other language tools, Resources and tools for building MT*, Budapest, pp. 41–49.
- Stokoe, W. C. (1960). *Sign Language Structure: An outline of the visual communication system of the American deaf*, 2nd printing, Burtonsville, MD (1993): Linstok Press.
- Veale, T., Conway, A. and Collins, B. (1998) The challenges of cross-modal translation: English to Sign Language translation in the Zardo system. *Machine Translation* 13, pp. 81–106.

Elicitation methods in the DGS (German Sign Language) Corpus Project

**Rie Nishio, Sung-Eun Hong, Susanne König, Reiner Konrad,
Gabriele Langer, Thomas Hanke, Christian Rathmann**

University of Hamburg

Institute of German Sign Language and Communication of the Deaf

Binderstr. 34, 20146 Hamburg, Germany

E-mail: {rie.nishio,sung-eun.hong,susanne.koenig,reiner.konrad,gabriele.langer,thomas.hanke,christian.rathmann}
@sign-lang.uni-hamburg.de

Abstract

During the first three years of the DGS Corpus project the main focus is on data collection. Before setting up the corpus design we conducted a survey to get an overview on the existing elicitation materials. The design of our data collection contains a variety of different stimuli and tasks with the special attention to free conversation, dialogues and monologues. To this effect, a range of possible discourse modes were considered: narration and renarration, discussion, report and description. The stimuli include pictures, picture stories, non-verbal film clips (e.g. cartoons and realistic film clips) and signed movies. In order to minimize the influence of the surrounding spoken/written language, written German is not used if possible. Introduction and explanation of each task is provided in DGS in form of movie clips. All tasks were tested in a pilot phase to examine their feasibility and reliability. Some of the tasks tested needed to go through several rounds of modifications while others did not work at all and thus were excluded from the data collection. In this paper, we not only present the tasks for elicitation and stimuli, but also describe their development process. We also discuss reasons why some stimuli were adopted from other projects while others had to be developed specifically for the purpose of our project.

1. Introduction

The DGS Corpus Project is a long-term project of the Academy of Sciences in Hamburg. It started in January 2009 and has two major aims: (i) to establish an extensive corpus of DGS and (ii) to develop a comprehensive dictionary of DGS-German based on the analysis of the corpus data.

In the first stage of the project, data of about 300 informants is collected at 12 sites throughout Germany. The corpus is designed to reflect everyday language of users of German Sign Language. The sample of informants is aimed to be balanced for sociolinguistic factors such as region, gender and age. Signers are always filmed in pairs and come for one elicitation session lasting for about 7 hours (including breaks). The target corpus size is a film length of 350-400 hours resulting in approximately 2.25 million tokens.

The purpose of the corpus is to document the use of DGS and also to provide material of and on Deaf culture and life. It will be a resource that can be used for a variety of research questions. About 50 hours of the material and its transcripts will be published for free access in the course of the project time. We expect that these materials will be interesting not only to researchers but also to the members of the Deaf community. In other words, we expect that the corpus not only becomes a valuable resource for linguistic research, but also a treasure given back to the Deaf community to which its members contributed themselves.

The corpus is compiled as a general resource for future research and is open as to what these questions might be. Therefore, it needs to consist of a large variety of discourse modes and grammatical structures as well as

various subject areas. As one of the project aims is to compile a general dictionary of DGS, the corpus should also provide enough material on the lexicon of DGS and its use.

In the following, we will discuss a survey we conducted on existing elicitation materials, describe the process of task development, and present the tasks along with the insights gained so far after completing the filming in two regions (Hamburg and Schleswig-Holstein).

2. Task development

2.1 Survey on existing elicitation materials

There are many different ways to collect language data from Deaf signers. There are visual stimuli such as pictures, photographs and movies, to name but a few. Since sign language researchers have used different elicitation materials for various research purposes since the early days of sign language research, our first aim was to get a comprehensive overview on the different kinds of elicitation methods. Although a number of elicitation materials have been shared among linguists, all of these stimuli are neither necessarily publicly available nor all researchers have published descriptions of their stimuli and elicitation procedures. A survey has been conducted within the sign linguistics community (Hong et al. 2009) to gather information on such materials. In the form of a questionnaire researchers were asked to give details on

- the form of the elicitation material used (pictures, animated cartoons, flash cards etc.),
- the content/subject matter of the elicitation material (topics of discussion, content of a

- picture story etc.),
- the research question,
- the specific task the informants were given.

In addition, the researchers were asked to give comments on the feasibility and reliability of their materials. The researchers were also asked if and from whom or for which project their material was adapted and if they were willing to share the stimuli with us.

On the basis of these questionnaires we were able to categorize the different kinds of elicitation materials in the following groups:

- language input (word lists of isolated words, single sentences in written language, written texts, signed videos),
- pictures (cartoons, single drawings, picture stories, photographs),
- motion pictures, movies, animations,
- topics for an open conversation or discussion (topical issues, fairy tales and fables),
- games,
- combination of pictures and words.

The analysis of the survey also allowed us to describe the advantages and disadvantages of the different stimuli. Furthermore, it became obvious that there are materials which are especially suitable for cross-linguistic studies because they have already been used for many spoken and/or signed languages, for example the picture book *Frog, Where Are You?* (Mayer, 1969), the so-called *Pear Story* (Chafe, 1980), the cartoon of *Tweety and Sylvester* (Warner Brothers, 1950) and the drawings from Zwitserlood (2003). Researchers' experience shows that elicitation materials cannot be adopted from abroad without taking cultural differences into account. Some linguists use the Aesop's Fables as a stimulus because these fables are well-known in many countries. This is not the case in Germany where most children grow up with Grimm's fairytales. But beyond the question whether Aesop's Fables or Grimm's fairytales are better known, linguists should always bear in mind that even such common stories might not be well-known within the Deaf community. The survey also indicates that a large number of stimuli of the same kind can be very tiring for the informants.

2.2 Adoption and development of tasks

We adopted *Frog, Where Are You?*, the *Pear Story* and the cartoon with *Tweety and Sylvester* in our tasks. The first two stimuli were originally used in spoken language studies (amongst others Berman & Slobin, 1994; Chafe, 1980) and were soon adopted by researchers in sign language studies (amongst others project "A Cross-linguistic Study of Sign Language Classifiers"). The cartoon *Tweety and Sylvester* is used for a cross-linguistic comparison of classifier constructions (project "A Cross-linguistic Study of Sign Language Classifiers").

Other existing materials could not be used or adapted because they did not meet our purpose. For example, the accessible stimuli for agreement verbs (e.g.

Hong, 2009) and negation (materials from the Centre for Sign Linguistics and Deaf Studies in Hong Kong) were designed to elicit isolated sentences. This is not the main focus of corpus building which should enable researchers to analyze signs and linguistic structures in a larger context of near-natural signing. For this reason, we developed new materials focussing on these phenomena (see 4.13 and 4.10).

Since one of the goals of our project is the compilation of a dictionary, the basic vocabulary was also in the centre of interest. None of the existing elicitation materials covers these needs. In order to collect the basic vocabulary which is not covered by the rest of the tasks, a task to cover as many subject areas as possible was developed (see 4.12).

Not only did we develop new tasks to elicit certain linguistic features, but we also ensured that different discourse modes are included in our corpus. For example, we created new tasks for eliciting negotiation and description of procedures (see 4.5 and 4.13).

In order to ensure that all informants would receive the same input, the instructions needed for each task were filmed in order to be presented to the informants on screen alongside with the materials. This also allowed us to provide different instructions to the two informants in settings where they had different roles in the task.

As for the adopted stimuli as well as new ones, we needed to deal with copyright issues. Two picture stories had to be excluded from our data collection because the publishers did not give us the permission to use the materials. Other publishers like the Bavarian Broadcasting (BR) and the Deaf Association in Berlin didn't have any objections and generously supported the project by providing us with materials.

3. Testing

3.1 Pre-tests

One step in the development of the various elicitation tasks was testing them in different stages of development to assess whether or not the tasks met our expectations. These tests were conducted by hearing researchers and student assistants with Deaf colleagues at the IDGS (Institute of German Sign Language and Communication of the Deaf) as informants.

After each test, the informants were asked if they felt comfortable with the task, understood the instructions clearly and if not, what they would suggest to improve them. In addition, they were asked if they considered this task suitable and feasible for potential Deaf informants. All tests were filmed and analysed to assess the following aspects of the tasks:

- Do the informants feel comfortable with the task?
- Do the informants understand the instruction movies? Is all necessary information given?
- Do the informants understand the stimulus material? Do they see what we want them to see?

- How much time does it take the informants to complete each task?
- How much signed output do the informants produce in each task?
- Do the informants produce the expected kind of language output (reliability)?

The tests revealed that in some cases the first versions of the instruction movies were not properly understood. This led to several rounds of revisions and re-testing before the final version was ready.

In some tasks the pre-test showed that an instruction movie alone was not sufficient for the informants. One of the aspects with which they had difficulties is the reference. The signer in the instruction movie addresses the informant directly by pointing forward and refers to the second informant by pointing behind his back due to the seating arrangements in the studio (see Hanke et al., this volume). Although the references established in the signed instructions matched the real elicitation setting, the informants did not understand the use of space in the instruction movie immediately. For this reason, the moderator, the fieldworker leading the session, now introduces the reference system at the very beginning of the session.

The pre-tests also made us aware that some informants tended to sign towards the moderator instead of signing to their dialogue partner. The moderators now get special training to avoid such situations.

Additionally, the stimulus material itself was edited. The font size of written words within the stimuli was enlarged and some pictures were replaced with better-known pictures, because the informants didn't grasp the picture's intention. One task, in which the informants are asked to describe the characters of the figures in an animated movie, had to be dropped since the informants tended to retell the story and had difficulties in describing only the characters of the figures.

As for all picture stories the pre-tests revealed that it is necessary to hide the stimuli when the informants is signing. Otherwise the signer would keep looking at the picture story instead of looking at his or her dialogue partner.

After the first testing period, the tasks were selected and put together in a reasonable sequence to get a session time of 5:30 hours with additional 1:30 hours for three breaks.

3.2 Final tests

Prior to the first elicitation session, we conducted two more or less complete test sessions each lasting a whole day. In the first session Deaf student assistants were recruited as informants and in another session two Deaf persons not affiliated with the IDGS were invited. The contact person in charge of the Hamburg area moderated both test sessions. The material and instruction movies were presented using SessionDirector (see Hanke et al., this volume) for the first time.

The major aim of these complete test sessions was

to simulate an elicitation session in a situation that was as close to the real studio setting as possible. The first session took place in a seminar room, but the second one could be held in the studio newly set up. In addition to the goals in the pre-tests, we also looked at the further aspects:

- How long does each task take, now embedded in the whole session?
- How long does the whole elicitation session take?
- Are the breaks at the right positions? How stressful is the session for the participants?
- Does the order of the tasks work? Do they influence each other in a positive or a negative way?
- Do interactions between the moderator and the informants work smoothly?
- Does SessionDirector work as expected in presenting the tasks and the stimuli? Do the informants know what to do when?
- Are Deaf people with different educational backgrounds able to cope with the tasks?

One result from the test sessions was the observation that the tasks took less time than in pre-tests and provided less material than expected. In the pre-tests the informants took much more time to complete each task. This may be an effect produced by the fact that the Deaf colleagues who served as informants in the pre-tests were used to signing in front of the camera, knew that they were expected to produce much signing and were therefore very cooperative. Another reason may be that in single tests the informants focus more on the given task while the participants in complete sessions knowing that the session contains many tasks and lasts for more than six hours focus more on completing the tasks than to linger on them. Here the results of the complete test sessions showed us that the moderator needs to be aware of the fact that the aim is not to complete the task as quickly as possible but to use the time and keep the informants on the subject to produce the expected amount of signed material.

As a result of the analysis of these two sessions we corrected the expected time for each task, modified tasks by adding subtasks and stimuli, changed the order of the tasks for the sake of balanced breaks, and added two extra tasks alongside the existing optional tasks to make the time management more flexible. We further refined the instructions to the moderator which are communicated in a written manual as well as in special training sessions.

4. Tasks

After the moderator has clarified questions concerning the consent form and checked on the questionnaire for the metadata collection with each informant, the moderator and the two informants take a seat in the studio to start the session.

4.1 Sign names

In the first task they are asked to show their sign names and to explain where these names come from. The goal of the task is to collect name signs with their origin as a part of Deaf culture. The task also aims at warming up the informants and introducing them to each other. We decided not to ask for their fingerspelled names of the informants (though they may present them, if they want), because some older Deaf are not familiar with fingerspelling. The whole task is completed in the average time of two and a half minutes.

4.2 Jokes

Prior to the elicitation session, each informant is asked to prepare one joke to present to the other informant on the day of filming. We adopted the idea of having one task for a prepared signing and its position at the beginning of the session from the Auslan Archive and Corpus Project. The task also helps the informants to warm up and to make them feel confident by signing something they are already familiar with. Furthermore, we expect that some of the informants tell a Deaf joke, which is part of the Deaf culture. Depending on the length of performances by both informants, the task takes between 2 and 7 minutes.

4.3 Experience of Deaf individuals

The moderator asks both informants questions on their experience from Deaf schools, residential schools, Deaf retirement homes, Deaf sports clubs, associations of the Deaf and so on to make them tell stories from their own lives. In this task no instruction movie is presented but instead the moderator needs to prepare questions in advance which fit the profile of the informant using the metadata questionnaire. The task aims at documenting typical experience from Deaf lives in form of narratives. We expect a lively and spontaneous talk as informants are supposed to tell their own experience. For this task the moderator is explicitly instructed to exploit the time slot of 20 minutes fully.

4.4 Movie and picture retellings

Informants look at either a picture story or a movie clip which they are asked to retell to the other informant. We paired four stimuli in two sets, so that one quarter of the informants performs each stimulus. Three of the stimuli are those which have been used in eliciting retellings in various languages: a picture story *Frog, Where Are You?* (Mayer, 1969), a movie clip with cartoon characters *Tweety and Sylvester* (Warner Brothers, 1950) and the so-called pear film or *Pear Story* (Chafe, 1980). The goal of using these stimuli is to supply materials for cross-linguistic research. The fourth stimulus is a comical sketch titled *Haushaltshilfe* (Housekeeper) broadcasted in the German TV program by and for the Deaf “Sehen statt Hören” (Bavarian Broadcasting, 2006). This is the only stimulus with DGS signing as an input in the whole elicitation session. (The exception is the

stimulus in an additional task, re-telling of the story on a fire alarm, see 4.18.) Both *Frog, Where Are You* and *Tweety and Sylvester* are presented twice. In the second run the story is divided into several groups of pictures / several movie clips and after each section the informant retells the respective part of the story. For our purpose some stimuli are presented in a slightly different form from the original. The *Pear Story* contains background sounds (but no verbalization), but it is played without sound. The broadcasted version of *Haushaltshilfe* is accompanied by German subtitles, but we use a version without subtitles, which the broadcasting company kindly provided. In the pilot phase Deaf informants pointed out that Deaf informants might get uneasy seeing signs in written English in *Tweety and Sylvester*, for which reason we considered adding German subtitles. However, we dropped the idea because the English signs did not have German counterparts and they did not play an important role in the story either. Rather, we decided to instruct the moderator to tell the informants to ignore the English signs. For the whole task the moderator is also explicitly instructed to turn the monitor black before the informant starts signing so that the informant doesn't look at the stimulus. This is important because the material then can be used in studies in which eye-gaze plays an important role. Since our experience in the final tests showed that the moderator sometimes forgets to do this, we adjusted the session directing software in a way that the monitor automatically turns black after 20 seconds in such cases. The pair of *Frog, Where Are You?* and *Tweety and Sylvester* takes 27 minutes on average to complete, *Pear Story* and *Haushaltshilfe* 17 minutes on average.

4.5 Calendar task

Informants are shown a one-week calendar with fictive appointments and are instructed to arrange two meetings of two hours respectively to prepare a surprise for the wedding party of a mutual friend. They are also told explicitly to talk about their other activities in the week during their negotiation. Target vocabularies of this task are days of the week, time terms and various common activities such as seeing the doctor, going on vacation, being at work, going to the movies and the theater, sports activities, having a plumber at home and so on. This is the only task in which some kind of role-play is required. The target discourse type is a dialogue with a special focus on planning and negotiation. We created two sets of calendars with different layouts, one with seven days side by side and the time flowing from top to bottom, like a timetable, and one with two pages for one week, Monday to Thursday being on the left and Friday to Sunday on the right page. In the pre-test, Deaf informants found the former more comfortable to look at. Nevertheless we kept both versions, because we realized that the Deaf informants to whom the former one was shown used vertical timelines in their signing which might have derived from the specific layout of the elicitation material. The task is completed in an average

time of 9 minutes.

4.6 Discussion

Informants are confronted with four controversial statements from which they are to choose one to discuss. The topics include both Deaf issues (e.g. cochlea implants, mainstreaming of the Deaf) and general issues (e.g. smoking bans). The goal of this task is to get the informants engaged in a lively and emotional discussion in which they hopefully don't think about their language use. We prepared two sets of topics, each of which is shown in every other session. The informants in the pilot phase mentioned their concern as to a high cognitive demand on informants as many of them are not used to reflecting on social issues or defending their own opinion. This makes the role of the moderator crucial who is supposed to put questions to support the informants to carry on their discussion. Our experience so far shows that they fill the slot of around 20 minutes. In some cases the moderator even needed to cut off the discussion to move on to the next task.

4.7 Free conversation

Following the topic discussion the moderator gives an instruction to the informants that they can now talk about anything they like while he or she leaves the room and comes back after 15 minutes. For ethical and practical reasons, the moderator makes explicit that the task is to chat in an unobserved setting. We adopted this task and its position after the topic discussion from the Auslan Archive and Corpus Project in which they had positive experience (p.c. Trevor Johnston July 16, 2009). In the DGS corpus project, the topics so far are the elicitation session itself, club activities (Deaf club, nine-pin club), family members and their hearing status, friends, communication and work.

4.8 Elicitation of isolated signs

Although our elicitation sessions mainly aim at filming monologues and dialogues, we have one task for eliciting isolated signs in order to document (regional) variation. In the first part of the task informants take turns at looking at German terms with or without an illustrating picture and are asked to sign it in DGS. Additionally, they are also asked to give one short example sentence of the sign. The choice of the 34 terms is based on previous experience from projects such as the sign language dictionaries of technical terms (e.g. Konrad et al., 2003). All of them have shown a wide regional variety in previous projects (e.g. bread, egg, water, man, woman, birthday, satisfied, mistake). In the second part, one informant is asked to sign the names of the 12 months and 4 seasons, and the other informant continues with 11 color terms for all of which a wide regional variety has also been observed. We intend to collect regional variation effectively and get some meta-linguistic discussions as one informant is free to comment on the sign or the example sentence of the other. The whole task is completed in 12 minutes on average.

4.9 Retelling of picture stories *Vater und Sohn*

In the final task of the morning session, each informant is asked to retell a simple picture story consisting of 5 to 6 pictures taken from the book *Vater und Sohn* (Father and Son) by Erich Ohser, a German cartoonist. We expect the informants to use constructed actions in their retellings. This is one of the optional tasks and can be skipped if other morning tasks took longer than expected. Our experience shows that the task takes the average time of 4 minutes.

4.10 Warning and prohibition signs

In the first task in the afternoon the informants look at warning and prohibition signs collected from different places of the world and are invited to discuss what they might possibly mean. In most cases the signs are unfamiliar to the informants and they need to guess. One practical aim of this task is to warm up the informants for the more demanding tasks in the afternoon. The scientific aim is to elicit negated sentences in a coherent context. The task turned out to be suitable for this purpose as our tests showed that the informants used negations in descriptions of the given signs, and occasionally, to express their disagreement to the other informant's suggestions. The task originally consisted of 12 warning and prohibition signs. Later, another 4 signs were added because the final tests showed that the discussions lasted slightly shorter than expected.

Our experience shows that the informants need an average time of 16 minutes to look at the instruction movie and discuss all of the 16 warning and prohibition signs.

4.11 What did you do when it happened?

In this task informants are asked to report what they did and/or felt when they heard about or experienced one of the shocking or moving events provided in the task. These include big historical moments (e.g. the moon landing, the fall of the Berlin Wall), significant soccer games in World Cups, catastrophes (e.g. the Indian Ocean Tsunami, the nuclear accident in Chernobyl), attacks (9/11, Kennedy assassination) and the death of famous figures such as Princess Diana. One of the topics is Deaf-specific, being the unexpected death of Gunter Trube, a widely recognized Deaf performer, an event which was a great shock to the German Deaf community. In addition to the signed description, well-known pictures of the events are provided which should evoke memories. The aim of the task is to encourage the informants to talk lively, in monologues (personal experience narratives) and/or in dialogues (further exchanges and discussions). The task also aims at documenting the way how Deaf people, who used to have limited access to information, learned about the news or experienced the events and how they processed them for themselves. In tests and in the elicitation sessions we indeed observed informants often mentioning TV news from which they had to guess what was going on. In order to cover various topics but not to

overwhelm the informants, we prepared two sets of stimuli, which are to be used in every other session. Each informant is asked to choose one out of 6 topics (or alternatively the informants choose two together). In the pilot phase and in the first elicitation sessions we got the feedback that younger informants were irritated by seeing not only recent events but also events, which eventually predated their birth. After a long discussion on whether to make specific sets for young informants, we decided not to make this age distinction in order not to reduce the flexibility of the setup should it become necessary to replace an informant (having fallen sick, for example) at short notice by someone else potentially from another age group. The task lasts 20 minutes on average.

4.12 Subject areas

This task is designed to initiate a conversation about at least two different topics. The aim is to get a solid basis for the selection of basic vocabulary in DGS. Therefore we classified every-day conversation into 25 subject areas (e.g. work and profession; energy and environment; family and relatives; ceremony, celebration and party; emotions and feelings; clothing and fashion; communication; partnership, relationship, love and sexuality; school and education; sports and games; travel). This classification takes former studies on basic vocabulary of written and spoken German into consideration (Plickat, 1980; Pfeffer, 1984) as well as actual lexicographic work on slang in spoken German (cf. Wippermann, 2009; and the corresponding website <http://szenesprachenwiki.de/>).

Each subject area is presented as a written German phrase with 4 to 8 photographs or drawings to complement the written input and to stimulate the informants' associations (figure 1).



Figure 1: Subject area work and profession

Due to the fact that we have at least 8 pairs of informants in each of the 12 locations, we prepared 8 different sets consisting of 4 subject areas each (some subjects appear in more than one set). To each pair of informants one set is presented. They are shown 4 slides with subject

name(s) and illustrations and a final slide, which summarizes the four subjects (with name(s) and at most 6 illustrations). The informants are to choose two subject areas. If they do not come up with anything to talk about, the moderator asks questions prepared by us for each subject area in order to initiate a conversation (e.g. "What do you find good about your job?", "Is there any law that is especially important to the Deaf?", "What can each of us do for a clean environment?"). If the informants are well ahead of time, one more subject (different from the suggested ones) is shown for further discussion. The task takes an average time of 32 minutes.

4.13 Combined tasks

This task is a combined task: one informant is supposed to perform the task description of procedures, the other one is supposed to retell a picture story. **Description of procedures:** The informant is asked to choose one familiar activity familiar to him/her from a set of 8. Each activity suggested consists of a sequence of actions (e.g. making jam, changing a car tire, decorating a Christmas tree). The target text types are step-by-step description and explanation. Furthermore, we aim at eliciting phrases to structure a text describing sequences of actions. We prepared two sets, each of which is presented in every other session, so that 16 activities are covered. However, if informants are not familiar with any of the suggested activities, they are free to describe any activity of their choice. **Retelling of a picture story** *Travel Story:* The informant looks at a picture story about a tour guide and participants who have to overcome several difficulties (figure 2).



Figure 2: Scene from the travel story

In the second run, the informant sees several pictures at a time and is asked to retell it to the other informant. As in the movie and picture retellings (see 4.4) the moderator is instructed to switch the monitor into black before the informant starts signing (or it turns black automatically after 20 seconds). The aim of the task is to elicit various ways of the use of space for directionality and plurality.

We created the picture story consisting of 17 scenes specifically for our purpose, because the survey mentioned above had shown that there were no suitable stimuli available for eliciting the target signs in the framework of a narration. The combined task, which consists of description of procedures and picture retelling takes 17 minutes on average.

4.14 Regional specialities

Informants are asked to talk with each other about specialities in the region they live in. The corpus design demands both informants living in the same region and having lived there at least for 10 years. Possible topics range from festivals of the region, popular tourist destinations, typical activities, famous figures, prominent landscapes, traditions and customs, typical products from the region to culinary specialities. The aim of the task is to collect signs for names of places, famous festivals and so on. The target text type is a discourse. We originally intended to elicit a planning discourse by asking the informants to produce a signed presentation on the region. The informants then would talk to each other about how to organize and prepare the signing output. We dropped this idea because most people (also hearing people) are not used to talk on a meta-linguistic level. The task lasts for about 20 minutes.

4.15 Retelling of a movie *Signs*

Both informants watch a five-minute movie and are asked to talk about it. The instruction is kept vague on purpose to avoid constraints on the conversation. What is special about the movie is the fact that there is no talking. The two protagonists communicate by showing each other written English words on a piece of paper. The end of the movie leaves it to the viewer to decide if the female protagonist is Deaf or not. We expect signs expressing love and feelings as well as assumptions. To make sure that the informants understand the written words in English, we added German subtitles. This task is optional and takes an average time of 8 minutes.

4.16 New vs. old signs

Informants are invited to report signs which are different between young and old generations. One goal is to capture sociolinguistic variance which is not covered in the other tasks. A further aim is to elicit a meta-linguistic discourse. In spite of the usefulness of the material we decided this task to be optional because during a pre-test we observed some discomfort among the informants who had difficulties in listing up such signs spontaneously. The task lasts 7 minutes on average.

We positioned two optional tasks, retelling of a movie *Signs* and new vs. old signs, near the end of the session to make the time management as flexible as possible.

4.17 Deaf events

The elicitation session ends with a Deaf-specific task in

which each informant is asked to talk about one Deaf event in which he or she took part. In order to call various Deaf events to mind, German names of the events and related visual materials (e.g. posters and pictures) are presented (figure 3). The topics range from national events such as culture festivals of the Deaf, sign language theatre festivals and sports festivals of the Deaf to international events such as Deaflympics and Deaf Ways. If the informant did not attend any of those events, he or she is free to choose any other event. The goal of the task is to document Deaf culture and to induce personal narratives and engaged conversations. The task takes an average time of 21 minutes.



Figure 3: Deaflympics

After this final task, the session ends with a closing conversation in which the informants are asked for feedback concerning the elicitation session itself.

4.18 Additional tasks

The moderator can include two additional tasks if the planned session time is not reached. One task is the retelling of a signed story about a fire alarm in a hotel and the other task is a route description based on a city map. Both of the tasks were adopted from the Dicta-Sign project (see Matthes et al., this volume). If the moderator decides to apply one or both of these tasks, they are inserted before the task "Deaf events" since we want our elicitation sessions to end with a Deaf-specific topic.

5. Conclusions

Having conducted about 20 elicitation sessions so far, the tasks and the elicitation session as a whole seem to work as expected. Due to the intensive pilot phase, in which many aspects could be reflected and improved, the stimuli achieve their intended purpose. Although the session lasts 7 hours including three breaks, the variety of topics and the diversity of task types seem to help the informants to work concentrated during the whole session. The feedback received so far from the moderators and the informants shows that the participants find most of the tasks interesting and entertaining. Thanks to the commitment of the

moderators and the motivation of the Deaf informants, the data collection started successfully. This provides a base for an extensive and valuable corpus, which will not only serve for future research, but also document the language and culture of the Deaf.

6. Acknowledgments

The research leading to these results has received funding from the German Academies of Science programme.

7. References

- Bavarian Broadcasting (2006): Sehen statt Hören. 1291. Broadcast on Oct. 7th, 2006.
- Berman, R.A., Slobin, D.I. in collaboration with Aksu-Koç, A.A. et al. (1994): *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, NJ: Lawrence Erlbaum.
- Chafe, W.L. (ed.) (1980): *The Pear Stories. Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Norwood, New Jersey: ABLEX.
- Hong, S.-E. (2009): *Eine empirische Untersuchung zu Kongruenzverben in der Koreanischen Gebärdensprache*. Seedorf: Signum Verlag.
- Hong, S.-E., Hanke, T., König, S., Konrad, R., Langer, G., Rathmann, C. (2009): Elicitation materials and their use in sign language linguistics. Poster presented at the Workshop “Sign Language Corpora: Linguistic Issues” in London, July 24-25, 2009.
- Konrad, R. et al. (2003): *Fachgebärdenlexikon Sozialarbeit/Sozialpädagogik*. Seedorf: Signum.
- Mayer, M. (1969): *Frog, Where Are You?* New York: Dial Books for Young Readers.
- Pfeffer, J.A., Lohnes, W.F.W. (eds.) (1984): *Grunddeutsch. Texte zur gesprochenen Gegenwartssprache. Textkorpora I. Einführungs- und Registerband*. Tübingen: Niemeyer.
- Plickat, H.-H. (1980): *Deutscher Grundwortschatz. Wortlisten und Wortgruppen für Rechtschreibunterricht und Förderkurse*. Weinheim und Basel: Beltz Verlag.
- Warner Brothers (1950): Canary Row. Broadcast on Oct. 7th, 1950.
- Wippermann, P. (ed.) (2009): *Duden – das neue Wörterbuch der Szenesprachen*. Mannheim: Dudenverlag.
- Zwitserslood, I. (2003): *Classifying Hand Configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands)*. Utrecht: LOT (<http://www.lotpublications.nl/index3.html>).

Glossing a multi-purpose sign language corpus

Ellen Ormel¹, Onno Crasborn¹, Els van der Kooij¹, Lianne van Dijken¹, Yassine Ellen Nauta¹,
Jens Forster² & Daniel Stein²

¹ Centre for Language Studies, Radboud University Nijmegen, PO box 9103, NL-6500 HD Nijmegen,
The Netherlands

² Human Language Technology and Pattern Recognition, RWTH Aachen University, Germany

E-mail: e.ormel@let.ru.nl, o.crasborn@let.ru.nl, e.van.der.kooij@let.ru.nl, l.vandijken@let.ru.nl, e.nauta@let.ru.nl,
forster@i6.informatik.rwth-aachen.de, stein@i6.informatik.rwth-aachen.de

Abstract

This paper describes the strategies that have been developed for creating consistent gloss annotations in the latest update to the Corpus NGT. Although the project aims to embrace the plea for ID-glosses in Johnston (2008), there is no reference lexicon that could be used in the creation of the annotations. An idiosyncratic strategy was developed that involved the creation of a temporary ‘glossing lexicon’, which includes conventions for distinguishing regional and other variants, true and apparent homonymy, and other difficulties that are specifically related to the glossing of two-handed simultaneous constructions on different tiers.

1. Introduction

Over the past years, various initiatives in the area of signed language annotation have been undertaken, but in the area of sign language glossing, no clear standards have been developed (Schembri & Crasborn, this volume). To some extent, researchers lean towards the general principles of the Leipzig Glossing Rules¹, but these do not specifically mention sign language data and the concomitant challenges. An important contribution to the discussion has been Johnston’s (2008) emphasis on the use on ‘ID-glosses’: identical forms should be consistently glossed, and variant forms should receive distinctive glosses.

Work on corpus construction, including the creation of annotations, has recently been increasing and is currently carried out for different sign languages other than Sign Language of the Netherlands (NGT), for example, for Auslan (e.g., Johnston, 2008; Johnston, 2009; Johnston, Vermeerbergen, Schembri, & Leeson, 2007), British Sign Language (BSL, e.g., Schembri, 2008), and German Sign Language (DGS, e.g. Hanke, 2002; Hanke, Konrad, & Schwarz, 2001).

Machine processing of signed languages has become an active research field as well, testified by the LREC workshop series. In order to facilitate machine processing of sign language corpora, several points need to be considered. In the present paper, we describe some of the adaptations in the Corpus NGT in order to facilitate machine processing.

A specific problem in the creation of the Corpus NGT was the absence of a lexicon with unique lemmata and variants that could be referred to. The dictionaries that have been published in the Netherlands are fragmented in focussing either on basic lexicon or on specific topics. In the last ten years, dictionary products have explicitly excluded variation with the aim of promoting standardisation of the lexicon (Schermer, 2003; Crasborn & Bloem, 2009; Crasborn & de Wit, 2005).

This paper will discuss the process of finding gloss conventions for the Corpus NGT that on the one hand function like ID-glosses, and on the other hand can be consistently created in the absence of a reference lexicon.

2. First release of the glossing conventions of the Corpus NGT

2.1 The Corpus NGT

The first release of the Corpus NGT in 2008 was created in a two-year project funded by the Netherlands Organisation for Scientific Research (NWO, grant no. 380-70-008), aimed at collecting a set of data from deaf signers using NGT (Crasborn, Zwitserlood & Ros, 2008, Crasborn & Zwitserlood, 2008). It has been completed in 2008, with the publication of the corpus on Internet.² The data consist of recordings with multiple synchronised video cameras, accompanied by gloss and translation annotations. All data are freely accessible to researchers and the general public. In each corpus video, a maximum of two subjects participated (S1 and S2). The left hand is glossed separately from the right hand.

All annotations were created in the ELAN software³ (see also Crasborn & Sloetjes 2008, 2010). This annotation tool allows multiple annotation layers (‘tiers’) to be time-aligned with several video files (Figure 1).

Every annotated file contains the following tiers: *GlosL S1*, *GlosR S1*, *GlosL S2*, and *GlosR S2*. These four tiers contain the glosses for the activities of the left hand (GlosL) and the right hand (GlosR) respectively, of the signer to the left (S1) and the signer to the right (S2). In signs made with two hands, the hands do not always move precisely simultaneously (Figure 2). Often, one hand stays in the final position of the sign, while the other hand starts articulating the next sign. Or one hand starts slightly earlier than the other hand. For each hand, the precise

¹ <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>.

² <http://www.ru.nl/corpusngtuk>

³ <http://www.lat-mpi.eu/tools/elan/>

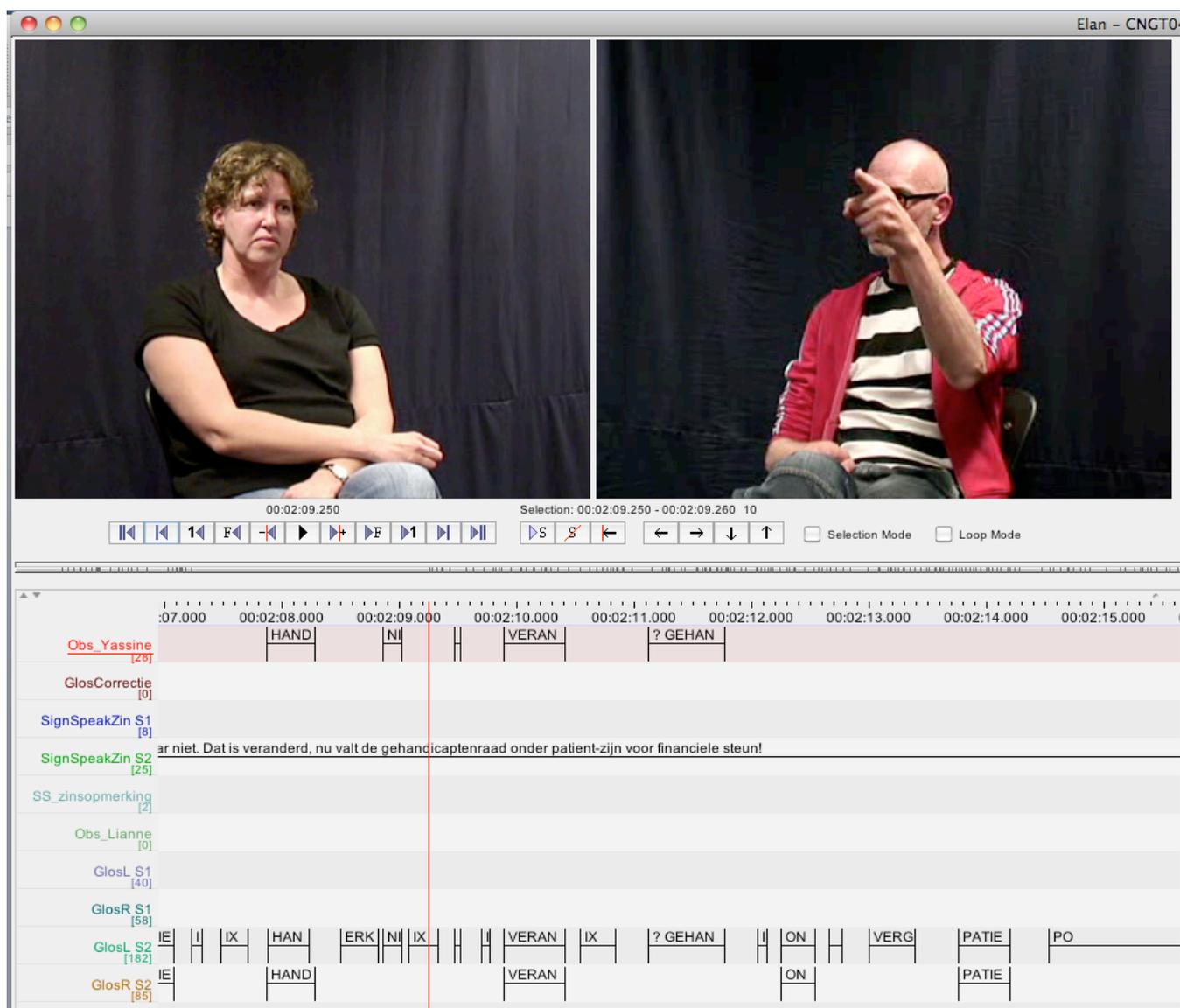


Figure 1. Multi-tiered annotation of multiple video files in ELAN

duration of the presence of a sign is shown in the gloss on the GlosL- or GlosR-tier.

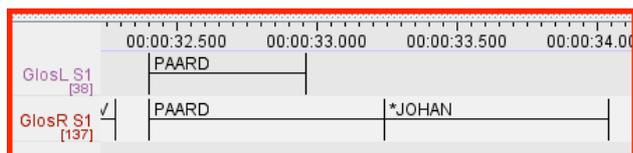


Figure 2. Time alignment of glosses per hand

2.2 Initial glossing conventions

The glosses in the annotation files in the *Corpus NGT* are

intended to indicate the exact start and end time of the signs, as well as to refer to a lexicon. Thus, the glosses in Dutch are not actual translations; in the ideal case they are pointers to lemmas in a lexicon. Because of the fact that there is no common orthography for sign language or a practical, commonly used phonetic notation system (Miller, 2001), Dutch words have been used as a reference, rather than first glossing in the language itself and subsequently translating the glosses to English or Dutch for accessibility, as is more commonly done for spoken languages. The Dutch glosses that are used approach (one of) the meaning(s) of the signs; however, the real meanings of the sign forms are described in the lexicon, not by the gloss. Exceptions to this rule are non-lexicalized forms that, in the gloss, are preceded by a @-character (see under #4 below).

Although it was our intention to use glosses referring to a lexicon, it was impossible to always consult the lexicons

of the Dutch Sign Centre (NGc) on DVD or on the internet, given the way the glosses were established and for reasons of efficiency. Because of this, the glosses in the first release contained many inconsistencies that users of the annotations need to be aware of. It is expected that many files contain a number of inconsistencies as well as interpretation differences and mistakes.

The glosses are primarily related to manual activity, not to body or facial activity, even though some lexical items include a specification of non-manual (mostly mouth) activities too. Non-lexical non-manual activity, as when the signer makes a manual sign accompanied by a head shake, are also not encoded by the gloss annotations: only the manual sign has been referred to in the gloss, not the negation expressed by the headshake.

3 Challenges

In order to improve machine search and machine processing of sign language corpora, several challenges need to be dealt with (Johnston, 2008). Most of these challenges stem from the fact that the glosses do not contain a transcription of the form of the language itself, but a pointer to a lemma in another language (Dutch, in this case).

The first challenge we have recently tackled in our NGT corpus concerns homonymy and polysemy. As for spoken languages, some signs have the same manual form, but do not share the meaning: homonyms, or do have the same manual form and have related but not identical meanings: polysemes. Lexicographers would define polysemes within a single dictionary lemma, while homonyms are treated in separate lemmata. In spoken English, the word ‘arm’ is an example of a homonym, which can refer to a limb, or it can be related to a weapon. In the first version of our sign language annotation conventions, homonyms and polysemes were ignored, as signs received a gloss based on the meaning of the manual part of the sign. In section 3, ‘revising the annotation convention’, we will discuss how homonyms and polysemes are currently processed. If homonym signs as well as polyseme signs would receive different glosses, an automatic recognition system would have severe difficulty grouping those signs that have the same forms.

A similar problem for recognition relates to the existence of sign variants: signs that have the same meaning, but a different form. Sometimes the same signers use these different signs as synonyms, but in addition there is some regional variation in the lexicon that the corpus recordings explicitly aimed to include. This type of variation was ignored in the initial release of the corpus as well, in that synonyms and/or regional variants simply received the same gloss.

Some additional challenges can be found in simultaneous constructions, whereby the left hand is articulating another sign than the right hand, which can even be one hand of a previous two handed sign (spreading) articulated simultaneously with a second sign. These types of special constructions are posing some real

challenges for machine recognition and translation systems, as those constructions convey a large range of creative combinations, which cannot be translated easily, let alone consistently. Classifier constructions pose an additional serious challenge for machine processing. Classifiers are non-lexical signs, which refer to a category of referents and their location and/or motion, and they too can be translated in multiple ways. For example, the sign for car can be used at first, and when referring to the car later in the discourse when it is driving across a hill for example, NGT signers use a flat hand, moving up and down a virtual hill.

4 Revising the glossing conventions

4.1 General revisions

Based on the need to adapt the glosses to facilitate machine-readability, a series of revisions have taken place. A thorough check of typos and spelling mistakes has taken place. At the same time, minor revisions of the annotation conventions such as notating ‘INDEX’ as ‘IX’ were implemented. Secondly, the time alignments between the video and the glosses were checked and adapted where necessary.

4.2 Umbrella glosses

An important addition to the first version of the conventions concerns signs that have identical manual forms, but differ in mouth pattern. These form a very frequent group of manual homonyms and polysemes. Some signs can have multiple meanings, depending on the context and whether or not a mouthing is used (Schermer, 1990, Crasborn et al., 2008, van de Sande & Crasborn, 2009).

We refer to part of those identical sign forms with related meanings (polysemes) by adding what we call an ‘umbrella gloss’ to the more specific gloss (examples will be discussed below). Signs that have an identical manual form can thus be labeled with a more general name, while keeping the information about the meaning in context.

The advantage of this approach is that during the annotation process, we do not have to make decisions on the exact status of the combinations between manual and non-manual activities: whether or not they form independent lemmata with fixed meanings is left to further research, but we facilitate further research by including a reference to both the manual form (by the umbrella gloss) and the contextual meaning (typically invoked by the action of the mouth). In a sense, this approach forms a midway between using phonetic transcription and foreign language labels, as it represents both the unique identification of the form as well as reference to the contextual meaning of the sign.

As the process of annotation continues, the number of umbrella glosses will increase. ‘AL’ (ALREADY in English) is an example of such an umbrella gloss. The label ‘AL’ is being used for various signs with an identical manual form, but with (somewhat) different

meanings. To further specify the sign, an addition is used, for example AL:GEWEEST (AL:BEEN), AL:GEHAD (AL:HAD) or AL:AF (AL:FINISHED). As with any sign language gloss, an (umbrella) gloss is not a translation of a sign; it remains a label attached to the sign. In the absence of a complete and accessible lexicon, it facilitates consistency in the glossing. In fact, an umbrella gloss can be chosen rather arbitrarily, as long as the label is used consistently. Below, two examples are listed for such signs that belong to an umbrella gloss: ALREADY and PROGRAMME. On the left is the more neutral gloss (the umbrella gloss), on the right the glosses that can be used when for example an accompanying mouth pattern adds to the meaning of the sign. When the sign has no accompanying mouth pattern, the more neutral term (or umbrella gloss) is used.

Umbrella gloss	Possible glosses if a sign has, for example, an accompanying mouthing.
AL (ALREADY)	AL:GEHAD (HAD) AL:GEWEEST (BEEN) AL:AF (COMPLETED)
PROGRAMMA (PROGRAMME)	PROGRAMMA:REGELS (RULES) PROGRAMMA:WETTEN (LAWS) PROGRAMMA:EISEN (DEMANDS) PROGRAMMA:PLAN (PLAN) PROGRAMMA:AGENDA (AGENDA)



Figure 3. Simultaneous constructions involving numeral buoys

In the online NGT lexicon⁴ these variants are not all listed as instances of a shared type; this is one of the reasons why it is hard to use an existing, fixed, lexicon for annotating a sign language corpus. The variation in the combinations of manual and non-manual forms can be used to further enhance existing lexicons.

4.3 True homonyms

A second part refers to identical forms as well, but instead of holding a shared meaning, these signs have highly distinct meanings: homonyms. An example from NGT is DOCTOR and BATTERY, which are both formed by the curved index and middle fingers touching the chin. Those types of homonyms will not share an umbrella gloss. For automatic sign recognition as well as for phonological and lexico-semantic research, it is crucial that such additional homonyms are listed separately as ‘true homonyms’, as separate from the polysemes that are joined by an umbrella gloss.

4.4 Regional variation in manual forms

Another addition to the conventions concerned sign translations that can have different sign forms, the so-called (regional and interpersonal) variants. It is important that different signs that have the same meaning (and therefore would receive the same gloss) but with a different sign form, can still be distinguished. The way to do this is adding a capital letter suffix to those glosses. A separate document is being made with different sign variants, for example: DOG-A and DOG-B.

4.5 Numeral constructions

Number signs were also in need of revised annotation conventions. Instead of glossing ‘counting hand’ for one hand, and ‘IX’ (pointing) for the other hand, we revised our gloss conventions so as to specify where exactly the dominant hand is pointing to. The gloss ‘IX’ is being used followed by a specification of the finger that is pointed at/indexed. Of this finger that is pointed at, only the first letter is glossed, e.g., IX:D (D for duim (in Dutch) = thumb) or IX:W (W for wijsvinger (in Dutch) = index finger). The non-dominant (counting) hand is specified for the number that is being realised, rather than merely stating ‘counting hand’. See the example in Figure 3.

4.6 Spatial variability

Some lexical signs can be performed in highly distinct manner, for example for direction verbs, such as ASK and VISIT. The glosses were adapted, such that different glosses are given, depending on the direction of the verbs. If a sign is directed from the signer towards another addressee, the gloss is composed of 1GLOSS, for example 1ASK, whereas if a sign is directed from an addressee towards the signers, the gloss is composed of GLOSS1, for example, ASK1. The number 1 refers to

⁴ <http://www.gebarententrum.nl>

the signer.

4.7 Sentences

Sentence boundaries are clearly needed for machine recognition and translation. However, although a series of boundary cues were found in past research, final conclusions on boundary markers have not been established thus far (Crasborn 2007, Nicodemus 2009). In order to facilitate sign language recognition, sign language sentence boundaries based on intuitive judgments were added to the annotations. Moreover, translations were provided for one topic in the corpus and boundary cues are examined, specifically designed for a European project; SignSpeak.

5. Conclusion

As for all language corpora, sign language corpora should be created in a systematically controlled and consistent way, which make machine searches and machines processing possible (Johnston 2008). This not only provides us the possibility to study linguistic properties in sign languages into much more depth and using much larger sign data sets than before, but, importantly, it has also resulted already in first steps towards automatic sign recognition and sign to text translations. In order to achieve this, we have revised the glossing conventions of the first release of the Corpus NGT in such a way that they consistently label specific forms, taking into account creative variations of which it is not clear whether they have been lexicalised or not. In this way, we also try to circumvent the absence of a well-accessible lexicon.

It will be clear from the discussion in this paper that we have aimed to create a workable solution that addresses the demands of both linguistic research and language technology. Further discussion on both details and principled choices is clearly necessary. A workshop of the *Sign Linguistics Corpora Network* in June 2010 is devoted to annotation, and will also take up the discussion on sign language glossing.⁵

5 References

- Crasborn, O., J. Mesch, D. Waters, A. Nonhebel, E. van der Kooij, B. Woll & B. Bergman (2007) Sharing sign language data online: experiences from the ECHO project. *International Journal of Corpus Linguistics* 12:535-562.
- O. Crasborn & T. Bloem (2009) Linguistic variation as a challenge for sign language interpreters and sign language interpreter education in the Netherlands. In Jemina Napier (ed.) *International perspectives on sign language interpreter education* (pp. 77-95). Washington DC: Gallaudet University Press.
- Crasborn, O. & H. Sloetjes (2008) Enhanced ELAN functionality for sign language corpora. In: O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood, eds. *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA. Pp. 39-43.
- Crasborn, O. & M. de Wit. 2005. Ethical implications of language standardisation for sign language interpreters. In *International perspectives on interpreting. Selected proceedings from the Supporting Deaf People online conferences 2001-2005*, edited by J. Mole. Brassinton: Direct Learn Services, pp. 112-119.
- Crasborn, O. & I. Zwitserlood (2008) The Corpus NGT: an online corpus for professionals and laymen, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd, eds. ELDA, Paris, pp 44-49.
- Crasborn, O., I. Zwitserlood & J. 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. <http://www.ru.nl/corpusngtuk>
- Hanke, T. (2002). iLex - A tool for sign language lexicography and corpus analysis. In: M. González Rodríguez, Manuel and C. Paz Suarez Araujo (eds.): *Proceedings of the third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, Spain; Vol. III. Paris: ELRA. 923-926.
- Hanke, T., Konrad, R., & A. Schwarz, S. (2001). GlossLexer – A multimedia lexical database for sign language dictionary compilation. *Sign Language and Linguistics* 4(1/2): 161–179.
- Johnston, T. (2008) Corpus linguistics and signed languages: No lemmata, no corpus. In: Crasborn, O., Hanke, T., Thoutenhoofd, E.D., Zwitserlood, I. and Efthimiou, E. eds. *Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the representation and processing of sign languages*. Paris: ELRA, pp. 82-87.
- Johnston, T. (2009) Guidelines for annotation of the video data in the Auslan Corpus. Ms., Macquarie University. http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf.
- Johnston, T., Vermeerbergen, M., Schembri, A., & Leeson, L. (2007). 'Real Data Are Messy': Considering Cross-Linguistic Analysis of Constituent Ordering in Australian Sign Language (Auslan), Vlaamse Gebarentaal (VGT), and Irish Sign Language (ISL). In P. Perniss, R. Pfau & M. Steinbach (Eds.), *Proceedings of the Workshop on Sign Languages: A Cross-Linguistic Perspective, Mainz, Germany, March 25-27, 2004*. Berlin: Mouton de Gruyter.
- Miller, C. (2001) Some reflections on the need for a common sign notation. *Sign Language & Linguistics* 4:11-28.
- Sande, I. van de & O. Crasborn (2009) Lexically bound mouth actions in Sign Language of the Netherlands. A comparison between different registers and age groups. *Linguistics in the Netherlands* 26: 78-90.

⁵ <http://www.ru.nl/slcn>

- Schembri, A. (2008) British Sign Language Corpus Project: Open access archives and the observer's paradox, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd, eds. ELDA, Paris, pp. 165-169.
- Schembri, A. & O. Crasborn (this volume).
- Schermer, Trude. 2003. From variant to standard. An overview of the standardization process of the lexicon of Sign Language of the Netherlands (SLN) over two decades. *Sign Language Studies* 3 (4): 96-113.

Video Analysis for Continuous Sign Language Recognition

Justus Piater, Thomas Hoyoux, Wei Du

INTELSIG Laboratory, EECS Department
University of Liège, Belgium
firstname.lastname@ULg.ac.be

Abstract

The recognition of continuous, natural signing is very challenging due to the multimodal nature of the visual cues (fingers, lips, facial expressions, body pose, etc.), as well as technical limitations such as spatial and temporal resolution and unreliable depth cues. On the other hand, signing gestures are designed to be robustly discernible. We therefore argue in favor of an integrative approach to sign language recognition that aims to extract sufficient aggregate information for robust sign language recognition, even if many of the individual cues are unreliable. Our strategy to implement such an integrated system currently rests on two modules, for which we will show initial results. The first module uses active appearance models for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow raise. The second module is dedicated to hand tracking using color and appearance. A third module will be concerned with tracking upper-body articulated pose, linking the face to the hands for increased overall robustness.

1. Introduction

Automated sign language recognition from video has been studied for at least about twenty years (Dorner, 1993). Most of this work has focused on the recognition of individual signs (Buehler et al., 2009; Cooper and Bowden, 2009; Yang et al., 2009), or placed heavy restrictions on grammar and vocabulary (Starner et al., 1998). The recognition of continuous, natural signing is very challenging, in terms of both video analysis and linguistics, due to the multimodal nature of the cues (fingers, lips, facial expressions, body pose), extralinguistic elements such as spatial references and pantomime, etc. These fundamental difficulties are joined by technical limitations such as spatial and temporal resolution and unreliable depth cues. On the other hand, serving communication, signing gestures are clearly designed to be robustly discernible. For example, while it is very difficult to estimate an articulated hand pose by matching a model to an image, relevant hand poses can be distinguished by appearance using supervised learning methods. Ambiguities in manual signs can often be resolved by integrating facial cues, etc. We therefore argue that an integrated approach to sign language recognition is required that combines the various visual and linguistic cues available using specialized, complementary techniques, aiming to extract sufficient aggregate information for robust sign language recognition, even if many of the individual cues may be unreliable at any given point in time (Dreuw et al., 2007).

Our own strategy to implement such an integrated system rests on two modules, for which we will show initial results. The first module uses active appearance models for detailed face tracking, allowing the quantification of facial expressions such as mouth and eye aperture and eyebrow raise. The second module is dedicated to hand tracking using appearance. It combines a discriminative method for selecting skin-colored regions with a generative method for characterizing hand configurations and locating images of hands in various articulated poses. This already permits a fairly robust estimation of hand trajectories.

2. Face Analysis

Facial expressions and head tilts play a very important role in sign language. Many manual signs are ambiguous in isolation, and need to be accompanied by appropriate facial expressions in order to convey a specific message. Moreover, facial expressions represent a continuous stream of supplementary information in any sign language communication, offering clarity and sensitivity to the viewer who actually looks more at the face than at the hands.

For computational purposes, facial parameters such as eye and mouth apertures can be inferred from the configuration of a set of relevant facial features such as the positions of fiducial points on eyelids and lips. Our face tracking system tracks such facial features using Active Appearance Models (Cootes et al., 2001)

Active Appearance Models (AAMs) are statistical generative models. Shape and texture variations of the human face as well as the correlations between them are learned from a set of example face images, on which corresponding “landmark” points have to be marked *priori* (including our facial feature points of interest). Fitting the AAM to a target image is done by finding the values of the parameters that minimize the difference between the synthesized model image and the target image using gradient descent. AAMs are very useful for our purposes because they offer a way to directly recover the structural parameters of a face and extract semantic content meaningful to the application. The complete framework of our face tracker is composed of (1) an offline part where we build the face model that contains all the facial appearance variation information as well as precomputed data for the step of fitting, and (2) an online part where we actually track facial features in real time using that model. Because the fitting method is a local search, we initialize the AAM using the face detector by Viola and Jones (2001). When the residual fitting error becomes high, we stop the tracking and come back to the detection step to reinitialize the model.

Fig. 1 shows feature extraction and expression quantification for four frames from a video sequence of the Corpus

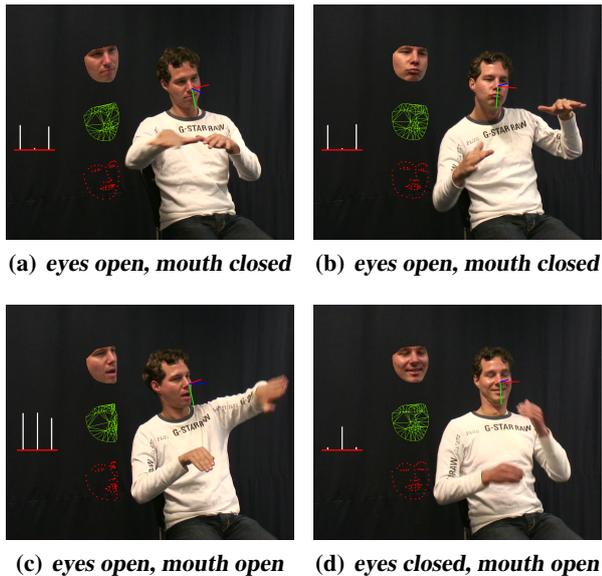


Figure 1: AAM fits – top-down: full model instance, meshed shape (green), plotted shape (red). Apertures (white lines) – left-right: left eye, mouth, right eye. Head pose: horizontal axis (red), vertical axis (green), depth axis (blue), the origin is the nose tip.

NGT, which is a collection of data from deaf signers using Sign Language of the Netherlands (Crasborn et al., 2008; Crasborn and Zwitserlood, 2008). Eye and mouth apertures shown here are quantified by the normalized area of the contours delimited by eye and mouth point features respectively. Head orientation is estimated using the POSIT algorithm, which gives the 3D pose of an object from a monocular view and the 3D structure of the object (DeMenthon and Davis, 1995).

Although AAMs constitute a powerful basis for building a face tracker, we need to apply refinements to the original formulation of this method to be able to robustly and accurately track facial features under the very uncontrolled conditions of the tracking scene in this project.

Often a signer’s face is partially occluded by the hands, and also self-occluded because of extreme off-plane head rotations. Local occlusions can lead the global model to degenerate and lose track of even non-occluded features, so we need to use particle filtering in combination with the AAM (Hamlaoui and Davoine, 2005).

It should also be pointed that large head rotations induce non-linearities in the 2D shape variation, which may not be robustly captured by the linear AAM model; a solution to this consists in using 2D+3D Active Appearance Models (Xiao et al., 2004) where the 3D structure of the face is learned and used to constrain the 2D AAM.

Finally, in this project we seek the most reliable (robust and accurate) AAM face model while preserving genericity, i.e. independence of the tracked person. In actual fact, we may accurately talk about independence of the video, since a person’s face can change over time, and since different imaging conditions can incur significantly different appearances of one person’s face. Since AAMs are statistical models of appearance, built with a learning procedure,

the genericity question is closely related to the choice of the training samples.

In AAM training, as in all learning tasks, one must carefully select the training examples, in quality as well as in quantity. An AAM is person specific if it is trained on examples of the face of one person only. If the examples are well chosen, the ability of the model to describe the face of this person in unseen situations is great. However, it will fail to accurately describe any other person. An AAM is generic if it is trained on examples of the face of several persons. In this case we can use the model to describe with good accuracy unseen faces of several persons, but with inferior accuracy compared to a person-specific model of the tracked person (Gross et al., 2005). Our research effort thus aims at finding ways to adapt a generic model to a specific face on the fly, combining the advantages of both methods while avoiding their drawbacks.

To illustrate the consequences of using specific or generic models, we built three AAMs on persons from the RWTH-Boston-104 database (Dreuw et al., 2007). We selected three videos: the first two videos show the same signer (a woman) but with significantly different appearances; the third video show a different signer (a man). The first model we built is specific to the first video of the female signer, and the second model is specific to the video of the male signer. The third model is built from images of the first and third videos; it is thus generic for two persons, or more precisely for two videos. Using each model in turn to track the face in each video, we compute the mean residual fitting error (i.e. the image difference between the best model instance and the target image, in the model reference frame) for each combination of a model and video. Tab. 1 shows the results thus obtained, and Fig. 2 shows some related sample images with the corresponding AAM fits, one for each model/video combination. Here, the specific models perform better than the generic model on the corresponding videos. Also, the model specific to video 1 poorly tracks video 2, even though it shows the same person.

AAM	video 1	video 2	video 3
specific to video 1	0.23	0.70	0.85
specific to video 3	1.60	1.10	0.12
generic (videos 1 and 3)	0.25	1.20	0.15

Table 1: Global performances of different models (specific and generic) presented with different data. The performance measure is the mean residual fitting error.

3. Hand Analysis

In sign language, hands convey a lot of information in different ways, including at least configurations, positions, trajectories, and instantaneous velocities of the two hands. These parameters are fairly difficult to extract robustly. In principle, hands are difficult to track, and their configurations (articulated pose) are difficult to estimate, because of their high number of degrees of freedom and their high level of self-occlusion, which give rise to an enormous variation of appearance and a high level of ambiguity. Thus, even if perfect image information were available, fitting an articu-

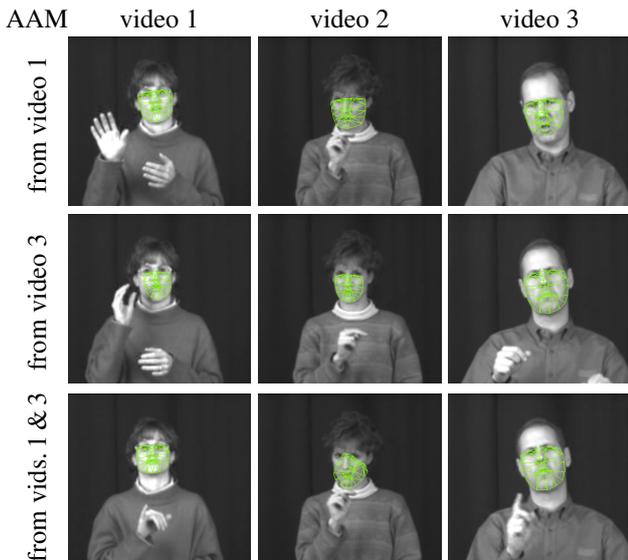


Figure 2: Sample images with AAM fits. Poor fits correspond to the inability of a model to interpret the data with which it is presented.

lated model of a human hand to image data is computationally hard.

These fundamental problems are exacerbated by technical issues. Most importantly, hands tend to move fast with respect to the frame rates and shutter times of typical video recording equipment, which results in substantial motion blur. Moreover, in typical recording settings, the structural determinants of the hands are small with respect to the pixel size, and imaging conditions are not optimized to enhance finger contrast. Consequently, the recovery of precise hand positions, let alone their articulated configurations, is very difficult in practice.

One promising path toward a solution rests on two methodological pillars, (1) discriminative machine learning methods that identify systematic predictors of specific hand-related parameters, and (2) the exploitation of redundancy. Our hand tracking system contains two steps that exploit these, skin-color region segmentation followed by PCA-based template matching.

For the segmentation of the skin regions, the popular graph-cut algorithm is adopted (Boykov et al., 2001). Graph cuts seek to minimize an energy function of the form

$$E = \sum_{p \in P} D_p(x_p) + \sum_{\{p,q\} \in N} V_{p,q}(x_p, x_q),$$

where D_p is called the data or unary term that measures how well label x_p fits pixel p given the observed data, and $V_{p,q}$ is called the smoothness or pairwise term that enforces smooth labeling among neighboring pixels.

For our skin segmentation problem, we incorporate two types of information in D_p . The first is a color likelihood based on histogram matching, and the second is a motion likelihood based on image differencing. The intuition behind is that hands of signers have distinct skin colors that are different from the background, and that the hands produce the most dramatic movement in sign language videos

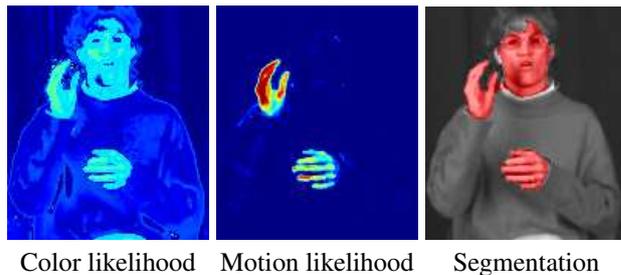


Figure 3: Color- and motion-based face and hand segmentation.



Figure 4: (Top) PCA bases of the left hand. (Bottom) PCA bases of the right hand.

(Fig. 3). For the smoothness term, we adopt the contrast-sensitive Potts model (Boykov and Jolly, 2001),

$$V_{p,q}(x_p, x_q) = \begin{cases} 0 & \text{if } x_p = x_q \\ \alpha + \beta \exp(-\frac{\|I_p - I_q\|^2}{\theta}) & \text{otherwise} \end{cases},$$

where I_p and I_q are the colour vectors of pixel p and q respectively. α , β , and θ are model parameters whose values are learned using training data. One example of skin segmentation is illustrated in Fig. 3.

After segmentation, we search hands in only the segmented skin regions using PCA based template matching (Ding et al., 2006). To this end, we collect training data from a few sign language videos and train PCA models for both the left and the right hands, shown in Fig. 4. Then, we randomly sample a number of hand candidates from skin regions, and match them with the PCA bases of the left and right hands. Thus, two matching scores are computed for each hand candidate reflecting the probability that the candidate is the left and the right hand. The hand model with the highest match score is most likely to be the hand being tracked in the current frame. However, we smooth hand trajectories over time by penalizing large motions between frames. This is currently done offline using dynamic-programming techniques (Godsill et al., 2001). Tracked hand regions and the corresponding PCA reconstructions are shown in Fig. 5.



Figure 5: The tracked hand regions, top row, and the PCA reconstructions, bottom row.

4. Conclusions

Automatic recognition of sign language requires the combined analysis of complementary modalities, including hand gestures, facial expressions, and body pose. We described our initial work on face and hand tracking. A third module for tracking upper-body articulated pose will be added at a later stage.

Both face and hand modules are as yet incomplete. Among the most important remaining problems of face analysis are the adaptation of generic face models to the face currently tracked to achieve genericity without sacrificing precision, and the estimation of gaze direction, which plays an important role in sign language interpretation.

Hand tracking is inherently difficult. Two fundamental problems are the difficulty of detecting hands in arbitrarily cluttered images, and the reliable distinction of left and right hands. To obtain reliable results, hand tracking should be informed by the configuration of the torso. To this end, hands are typically tracked in conjunction with the arms, which are further constrained by the positions of the shoulders with respect to the head (Buehler et al., 2008). Again, by themselves, arms are difficult to track because their appearance is usually very similar to the upper body of the tracked person; all there is to exploit are weak and ambiguous edge cues. However, combined with an articulated body model as well as face and hand tracking, reliable overall results can feasibly be obtained.

The principal remaining difficulty for upper-body tracking is the extreme variation of upper-body appearance between signers. This can be overcome e.g. by requiring an initial, instantaneous initialization from a canonical pose, which is used to bootstrap online learning of a discriminative appearance model for hands and arms. In addition, we are working on exploiting non-local motion cues to inform the hand tracker, increasing robustness in ambiguous situations such as low contrast and occlusions between hands and arms.

5. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007–2013 – Challenge 2 - Cognitive Systems, Interaction, Robotics – under grant agreement n° 231424-SignSpeak.

6. References

- Y. Boykov and M. Jolly. 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *International Conference on Computer Vision*, volume I, pages 105–112.
- Y. Boykov, O. Veksler, and R. Zabih. 2001. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239.
- P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. 2008. Long term arm and hand tracking for continuous sign language TV broadcasts. In *British Machine Vision Conference*.
- P. Buehler, M. Everingham, and A. Zisserman. 2009. Learning sign language by watching TV (using weakly aligned subtitles). In *Computer Vision and Pattern Recognition*.
- H. Cooper and R. Bowden. 2009. Learning Signs from Subtitles: A Weakly Supervised Approach to Sign Language Recognition. In *Computer Vision and Pattern Recognition*, pages 2568–2574.
- T. Cootes, G. Edwards, C. Taylor, et al. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.
- O. Crasborn and I. Zwitterlood. 2008. The Corpus NGT: an online corpus for professionals and laymen. In Crasborn, Hanke, Zwitterlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, pages 44–49, Paris. ELDA.
- O. Crasborn, I. Zwitterlood, and J. 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.
- D. DeMenthon and L. Davis. 1995. Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1):123–141.
- C. Ding, D. Zhou, X. He, and H. Zha. 2006. R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pages 281–288.
- B. Dörner. 1993. Hand shape identification and tracking for sign language interpretation. In *IJCAI Workshop on Looking at People*.
- P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, and H. Ney. 2007. Speech Recognition Techniques for a Sign Language Recognition System. In *Interspeech*, pages 2513–2516.
- S. Godsill, A. Doucet, and M. West. 2001. Maximum a posteriori sequence estimation using Monte Carlo particle filters. *Annals of the Institute of Statistical Mathematics*, 53(1):82–96.
- R. Gross, I. Matthews, and S. Baker. 2005. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093.
- S. Hamlaoui and F. Davoine. 2005. Facial action tracking using an AAM-based condensation approach. In *IEEE ICASSP*. Citeseer.
- T. Starner, J. Weaver, and A. Pentland. 1998. Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375.
- P. Viola and M. Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proc. IEEE CVPR 2001*.
- J. Xiao, S. Baker, I. Matthews, and T. Kanade. 2004. Real-time combined 2D+ 3D active appearance models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2. Citeseer.
- H.-D. Yang, S. Sclaroff, and S.-W. Lee. 2009. Sign Language Spotting with a Threshold Model Based on Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(7):1264–1277.

Data-Driven Sub-Units and Modeling Structure for Continuous Sign Language Recognition with Multiple-Cues

Vassilis Pitsikalis, Stavros Theodorakis and Petros Maragos

National Technical University of Athens, School of ECE, Athens 15773, Greece.
{vpitsik,sth,maragos}@cs.ntua.gr.

Abstract

We investigate the automatic phonetic modeling of sign language based on phonetic sub-units, which are data driven and without any prior phonetic information. Visual processing is based on a probabilistic skin color model and a framewise geodesic active contour segmentation; occlusions are handled by a forward-backward prediction component leading finally to simple and effective region-based visual features. For sign-language modeling we propose a modeling structure for data-driven sub-unit construction. This utilizes the cue that is considered crucial to *segment* the signal into parts; at the same time we also *classify* the segments by implicitly assigning labels of Dynamic or Static type. This segmentation and classification step disentangles *Dynamic* from *Static* parts and allows us to employ for each type of segment the *appropriate* cue, modeling and clustering approach. The constructed Dynamic segments are exploited at the model level via hidden Markov models (HMMs). The Static segments are exploited via k-means clustering. Each Dynamic or Static part, exploits the appropriate cue related to the movement. We propose that the movement cues are normalized in order to be translation and scale *invariant*. We apply the proposed modeling for further combination of the movement trajectory individual cues. The proposed approaches are evaluated in recognition experiments conducted on the continuous sign language corpus of Boston University (BU-400) showing promising preliminary results.

1. Introduction

Sign languages, i.e., languages that essentially convey information via visual patterns, commonly serve as an alternative or complementary mode of human communication. Visual patterns, as opposed to the audio ones used in the oral languages, are formed by hand shapes and manual or general body motion, lip movements and facial expressions. Their expressiveness facilitates human interaction and exchange of information not only in the existence of hearing-impaired people but also in situations where speech is impractical, e.g., in loud workspaces. However, efficient communication by these means is only feasible between specially trained interacting parties. In this context, automatic sign-to-text and text-to-sign translation can be viewed as the intermediate technological modules that can partially lift this restriction. Moreover automatic sign language recognition may have contributions across other areas as linguistics for the study of sign languages or for the semi-automated processing of corpora.

Early attempts on automatic Sign Language Recognition (SLR) were restricted to simple recognition tasks [Ong and Ranganath2005] similarly to cases of speech recognition a few decades ago. An informal correspondence of the word in spoken language is a sign unit, given that sign languages tend to be monosyllabic [Emmorey2002]. There are several metaphors between sign and speech recognition that allow for the exchange of methods between the two areas. However, there exist points of difference too. A diversity that also has practical effects concerns phonetic sub-units: There is not yet a well-defined unit equivalent to the phoneme in speech. Another difference concerns the multiple parallel cues that are articulated during sign language generation. In this paper, as far as the segmentation, modeling and recognition are concerned,

we focus on automatic data-driven modeling of sub-units without any phonological or linguistic information.

The field of SLR is certainly in the focus of quite intense research lately [Ong and Ranganath2005]. It is considered to be a multilevel problem and it poses significant challenges regarding visual processing and information stream modeling for recognition. [Vogler and Metaxas2003] broke down signs into their constituent sub-units using the basic ideas of the Movement-Hold model [Liddell and Johnson1989] and applied successfully the so-called Parallel HMMs. [Bauer and Kraiss2001], on the other hand worked also at the sub-unit level exploring a data-driven approach for modeling the intra-sign units. They cluster independent frames utilizing K-means. [Fang et al.2004] and [Han et al.2009] have also proposed approaches for data-driven sub-unit modeling. They employed clustering by considering segments and not only independent frames as [Bauer and Kraiss2001] at the feature level, taking advantage of the dynamics that are essential in sign language. Modeling at the sub-unit level provides a powerful method in order to increase the vocabulary size and deal with more realistic data conditions.

The main objective of the proposed modeling approach is the automatic segmentation and construction of data-driven sub-units: these sub-units are the intra-sign primitive segments that shall be reused to reconstruct signs that share similar articulation parameters. We are inspired by both perceptual and linguistic evidence [Emmorey2002, Liddell and Johnson1989] on the functionality of the multiple cues. Among all cues the ones that we heavily exploit next are based on the planar (2D) coordinates of the dominant hand's centroid, and some of its products. We shall refer to these features from now on as the *movement-position* cues. Besides, movement and position are among the main characteristics that describe a sign [Emmorey2002].

Based on simple movement, position measurements, we proceed on the automatic sub-unit modeling of sign lan-

This research work was supported by the EU under the research program Dictasign with grant FP7-ICT-3-231135.

guage at the model level, that refers to the modeling of intra-sign segments. This modeling involves the synergy of the multiple cues and the modeling structure that these cues are incorporated: 1) the partitioning of segments into dynamic or static with respect to their dynamics; we employ for each sign unit, a model based segmentation at the state level, that apart from the segmentation assigns also labels to the segments. 2) The modeling of the static or dynamic segments depending on the label that they were assigned in the previous modeling step. Each type of segment shall be modeled by the cues and the model that are more appropriate for each case. Given the segmented sign we are equipped with a prosperous initialization step to face appropriately the modeling the dynamic vs. static intra-sign segments. For the case of dynamic segments, our goal is to cluster not the independent frames as if they were in a common pool [Bauer and Kraiss2001], neither the feature frames sequences as segments themselves at the feature level [Fang et al.2004, Han et al.2009]. Instead, we propose to hierarchically cluster whole dynamic models (HMMs) [Smyth1997] based on a similarity measure among models via their parameters. Another point to stress is that the models are first normalized wrt. 1) the initial segment's position, for each segment, and 2) the maximum scale of the movement's trajectory. These normalization steps are crucial, since by incorporating them we end up modeling the actual movement data independently to the existing mixed scales or initial positions: this makes the models more compact, increases the training data per model, and reduces the total number of models required. For the case of static segments, the main measurement that characterizes them is the one of position, corresponding to more clear postures on which the velocity has been on average close to zero. We evaluate the proposed methods on real data from the Boston-University continuous American Sign Language corpus (BU400) [Dreuw et al.2008]. In the experiments we explore a variety of feature streams and evaluate the efficacy of the proposed modeling scheme in preliminary automatic recognition experiments showing promising results. These experiments investigate the efficacy of the employed features, as well the integration of the multiple movement-position cues.

2. Visual Processing of Sign Language

2.1. Segmentation and Tracking

For the segmentation of the video frames we are based on the Geodesic Active Regions (GAR) approach [Paragios and Deriche2002], as this has been adapted on previous work [Diamanti and Maragos2008] for sign language processing. The GAR are deformable 2D contours, which evolve to minimize an energy functional, designed to meet the needs of the segmentation process. The intensity image is partitioned into two separable regions, one being the union of the skin-colored regions, and the other consisting of the rest of the image pixels, referred to as background. We adapt the GAR model to introduce a new force for skin segmentation.

$$F_{color} = \log((P_s(\vec{x}))/P_b(\vec{x})) + cg(I) \quad (1)$$

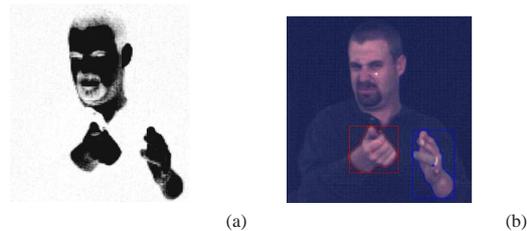


Figure 1: (a) Likelihood ratio per pixel belonging to skin or not, shown as a grayscale image. (b) Segmentation after employing GAR on the likelihood ratio map.

where I is the image, P_s , P_b denote the probability of a certain pixel \vec{x} belonging to the skin or background regions, respectively, and $g(I)$ is the edge-detection stopping function. To estimate the probabilities P_s and P_b we employ two probabilistic models to account for the skin and background color, respectively. After the estimation of P_s and P_b by taking their ratio we result with a measure of a pixel belonging to skin. The above likelihood ratio map is then used as a force in the GAR model enforcing the curve to converge eventually to the edges that separate the skin region from the background. The result of the hand detection that we use is shown in Fig. 1. Due to the dynamic nature of sign language articulation, the skin color regions of interest may occlude each other. For these cases we employ techniques in order to disambiguate occlusions such as linear forward-backward prediction and template matching.

2.2. Feature Summary

Employing the segmentation and tracking process, we extract features related to the position and the movement. More specifically using the fitted ellipses on each hand we extract the features related to these ellipses such as x, y centroid coordinates. In addition, we construct features which are products from the x, y coordinates of the hands' centroids, such as the velocity $vel(t) = [\dot{x}; \dot{y}]$, the acceleration $acc(t) = [\ddot{x}; \ddot{y}]$ and the instantaneous direction $dir(t) = [\dot{x}; \dot{y}]/(\dot{x}^2 + \dot{y}^2)^{1/2}$. For the scope of our current modeling and recognition we are using only the x, y coordinates of the dominant hand centroid using as reference point the centroid of the signer's head and its aforementioned products.

3. Continuous Sign Language Recognition

We tackle the issue of sub-unit probabilistic modeling in order to deal with continuous sign language recognition. We propose 1) the organization of the modeling in a tree-like modeling structure that employs on each modeling level the *appropriate feature(s)* with the appropriate modeling depending on the functionality of the features; 2) the normalization of the features that are modeled: We focus in this way on the actual underlying phenomena we wish to tackle and avoid from getting our modeling consumed on mixed factors; 3) the incorporation of the dynamics *at the model level* – and not at the feature level of separate frames or sequences of frames' level. We consider that it is both 1) the *modeling structure* and 2) the modeling with *normalization*, that are important as it is discussed next.

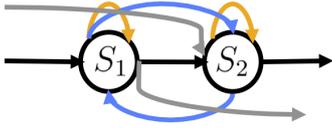


Figure 2: The 2-state HMM topology that is employed for segmentation and implicit classification of the segments.

3.1. Model-based Segmentation and Classification

Modeling the Velocity Cue Our goal is to separate the so called from now on, “dynamic” from the “static” parts w.r.t. movement. This is the level of segmentation and classification of the segmented parts of the signal: dynamic parts shall correspond to movements and static parts to non-movements. This approach is inspired by linguistic modeling [Liddell and Johnson1989] of “movements and holds”. We assume that movements correspond to high on average velocity, and non-movements to low relative velocity. Although the fuzziness of the ‘high’ and ‘low’ terms we appropriately incorporate them by adopting a suitable model-based approach. The feature that shall be utilized for this characterization is the *velocity*, whereas the *acceleration* could add further detail. The velocity feature vector consists of the dominant hand’s centroid velocity that is constructed as the norm of the coordinate derivatives. Our goal is met if we consider the HMM structure of two states, as shown in Fig. 2. This allows the entrance and the exit from both states and the full transition from each state to the other, since the dynamic or static parts may alternate one another and do not obey any grammar rule.

Gloss Specific Modeling Next, we create one model for each gloss that is trained using all realizations of the specific gloss. Each HMM gloss model models the velocity profile of the corresponding gloss. Each one of the HMM states results in modeling a single velocity level. Given the population of data from large portions of the training set, the two state levels correspond to a low- and a high-level of the corresponding feature, i.e. velocity. This is further understood if we observe the velocity distribution over the different realizations for a specific gloss in Fig. 4(a). After training each HMM we perform a Viterbi alignment for each realization given the gloss resulting to the most probable *segmentation points* at the state level *together* with the labels of the velocity profiles. An example of segmentation obtained for one instance of the sign “ADMIT” is depicted in Fig. 4(b) for the feature level, whereas Fig. 3 shows the actual frames of the segments (subsamped).

Automatic vs. Manual Segmentation One way to evaluate the proposed segmentation approach is by comparing its output with the corresponding manual annotation by experts. At this point we show the results of a preliminary such effort. Figure 5 presents both the automatic and manual annotation¹ for a realization of the sign “ADMIT”. For the automatic production of both segmentation points and the classification of the segments we make use of the veloc-

¹The manual annotation has been provided by Annelies Brafort at CNRS-LIMSI.



Figure 3: Segmentation using the velocity cue for one instance of the sign “ADMIT”. Each row corresponds to a different segment.

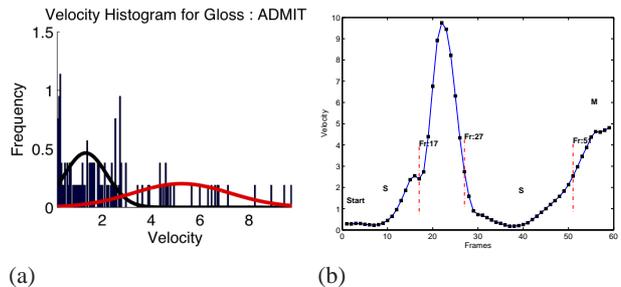


Figure 4: (a) Velocity distribution (histogram) superimposed with the fitted (b) Segmentation shown superimposed on the velocity profile for an instance of the sign ADMIT.

ity modeling providing two different labels. By comparing the results it seems that the automatic segmentation via the proposed approach is on average close to the manual segmentation points.

The proposed model-based approach provides various advantages: 1) we get not only the segmentation but also the result of a classification since we have encapsulated implicitly the dynamic and static characteristics into the states of the same model. 2) Another asset is that we don’t need to define any threshold manually (as other approaches for segmentation at the feature level), since these are handled inherently after setting the model parameters.

3.2. Modeling Dynamic Segments

We tackle next the issue of intra-sign sub-unit modeling at the HMM model level instead of the feature level. In this way we take advantage of the explicit dynamic modeling at the state level that the HMMs yield. Dynamic modeling is crucial for the modeling of movement. After all, HMMs have been employed successfully in other applications of sign language modeling too [Vogler and Metaxas2003]. Afterwards, a model level approach adds up a probabilistic viewpoint that can be further exploited, and finally fits well with the automatic recognition framework.

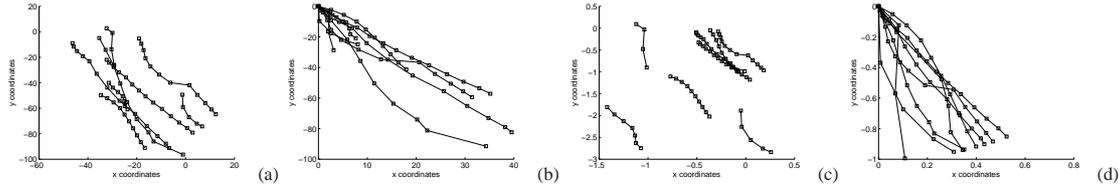


Figure 6: Trajectories of dynamic movements mapped onto the 2D signing space: (a) Without any normalization. (b) After normalization to the initial position. (c) After normalization to scale. (d) After normalization to both the initial position and scale.

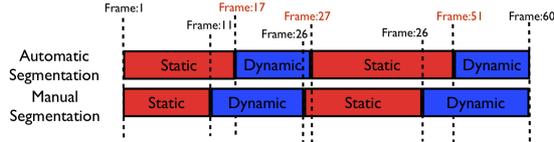


Figure 5: Automatic vs. Manual Segmentation and segments' classification for a realization of the sign "ADMIT".

3.2.1. Feature Normalization

Initial Position Our goal in this task is to model the dynamics of movement during the signs. The main feature for each dynamic segment is the *movement trajectory*. Each position sequence is initiated from the previous actual position that is arbitrary. The modeling of such features, leads to the consumption of the modeling effort due to the increased variance that the arbitrary initial positions of the movement trajectories introduce, so as to account for all different initial positions. This is encountered by normalizing the feature segments, each one with its corresponding initial position. This step results on the translation invariant movement modeling, i.e. independently to the initial position. An example of this normalization is presented in Fig. 6(a,b): we present the movement trajectories as they are mapped onto the initial 2D signing space before they are employed in the sub-unit construction process; we demonstrate the *same* trajectories with and without normalization. Moreover, normalization methods are well-known in the ASR community [Rabiner1989]. Another advantage of the normalization is the increase of the data requirements per model and at the same time we decrease the total number of models required.

Scale Similarly to the above, scale also affects the modeling of movement trajectories. Scale normalization of each movement results in scale invariant modeling, increase of data examples per model, end more efficient modeling with less models. At the same time, we do keep the scale parameter itself for further incorporation and modeling as a separate feature. An example of this normalization is shown in Fig. 6(a,c): the figure shows the same segments before they are employed in the sub-unit construction procedure with and without normalization. Finally Fig. 6(d) shows the same trajectories after both scale and initial position normalization. It shall be next more efficient to incorporate these normalized segments in the corresponding HMM models instead of the non-normalized, since they shall cap-

ture the actual dynamics independently to both the initial position (compare with Fig. 6(a,c)), and the maximum scale (compare with Fig. 6(b,c)).

3.2.2. HMM Clustering

We initialize the segments by first applying the segmentation procedures, as it has been described in the previous Section 3.1.. Given that the segments contain movement our goal is to cluster whole dynamic models (HMMs) [Smyth1997] that correspond to these movement trajectories. Clustering states at the model level has been employed successfully in ASR applications. Herein we cluster not just the states, but *whole* HMMs. Thus, we fit N 3-state HMMs, one for each individual sequence or segment S_i , $i = 1 \dots N$. Afterwards we use a similarity measure between pairs of HMM models H_k , $k = 1, 2$, by adopting among proposed approaches in the literature [Juang and Rabiner1985] that are based on the Kullback-Leibler divergence. Similarly we employ

$$D(H_1, H_2) = \sum_{O_i^{H_1}} \frac{1}{T_i} \log \frac{P(O_i^{H_1} | H_1, S_i^{H_1})}{P(O_i^{H_1} | H_2, S_i^{H_2})}$$

where $O_i^{H_k}$ corresponds to the observation sequences that have been generated from each of the H_k model, of length T_i and $\log P(O_i^{H_k} | H_k, S_i^{H_k})$ is the log probability of the observation given the HMM model and the optimum state sequence $S_i^{H_k}$, for $k = 1, 2$. The sequences used to compute the log probabilities are generatively constructed by each H_k model employing 20 sequences. The distance similarity matrix among all models is exploited via an agglomerative hierarchical clustering algorithm. We end up with the total likelihood of the specific clustering, given the number of clusters employed.

3.3. Dynamic Sub-Units for Each Feature

Next, we explore the modeling of features that are appropriate for dynamic segments modeling. The output of the clustering on the HMM level corresponds to a partition on the feature space. Each cluster in this partition is defined as a distinct sub-unit, presented next for different cases of features.

Movement Trajectories After the normalization steps each segment is modeling the plain normalized trajectory in the 2D planar signing space. We show in Fig. 7(b) indicative sub-units: these are clusters that have been constructed by the HMM hierarchical clustering at the model level, and are then mapped onto the 2D signing space. This mapping retains the sub-unit identity that is encoded by means

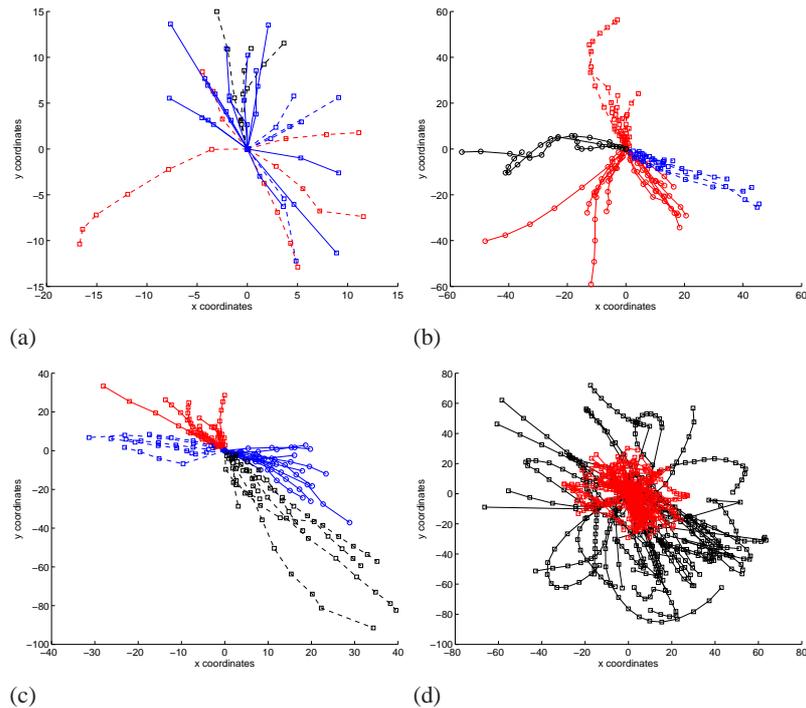


Figure 7: The trajectories for different sub-units as they have been mapped on the 2D signing space. With different color we represent different sub-units that correspond to the different clusters. (a) Trajectories of sub-units obtained using as feature the movement trajectories (P) *without* any normalization. (b) Trajectories of sub-units that incorporate both scale and initial position normalization. (c) Trajectories of sub-units that incorporate the Direction cue after normalization of the trajectories to the initial position. (d) Trajectories for two different sub-units that correspond to different scales.

of color in the presented figures. In Fig. 7(a) we show a case of sub-units as a result of clustering, but without the normalization steps. It is evident by comparing with the previous case (Fig. 7(b)) that the modeling is much looser since the models are consumed totally on the explanation of the different initial positions or scales. The non-normalized constructed sub-units as shown mapped on the original 2D signing space make it hard to understand what exactly each cluster represents. The clusters after normalization actually implicitly incorporate direction information. This is something expected as the modeling contains the direction information encapsulated with the geometry of the whole trajectory. As a matter of fact, each model's state from the first to the last explains points in the trajectory that have on average increasing distance from the segments initial position.

Scale We may have normalized with the scale of each trajectory, being the maximum distance of all points in a trajectory, but this information shall not be disregarded. It is modeled on its own in order to investigate how it affects the modeling. We show in Fig. 7(d) indicative sub-units: these are clusters that have been constructed by the clustering at the model level, and are afterwards mapped on the 2D signing space. This mapping retains the sub-unit identity or equivalently the cluster index that is encoded by means of color in the presented figures. The presented sub-units are presented to model trajectories entirely based on their scale independently to their direction.

Direction The sub-units constructed by the direction feature show similar results as the ones that model the nor-

malized trajectories. As expected each sub-unit consists of movements with similar on average direction over time. Figure 7(c) shows indicative examples of movements over the same or different clusters having similar on average or different directions respectively.

3.4. Dynamic Sub-Units for Multiple Features

In the previous section we presented the sub-unit construction for the dynamic segments using a single cue at each time for each sub-unit type. Thus we constructed sub-units that account for single different characteristics of a movement such as the direction, the scale or the movement trajectory. Next, we explore sub-unit construction for the dynamic segments by using for each sub-unit multiple cues. This extension is seamlessly incorporated given the discussed framework. As mentioned in Section 3.2.2. the sub-unit clustering is based on HMMs. In order to account for multiple features during sub-unit construction we employ a multi-stream HMM instead of one simple single-stream HMM. More specifically by incorporating both direction and scale into a multi-stream HMM we create multiple-cue sub-units that model movements based jointly on their direction and their scale. This sub-unit construction is shown via the corresponding trajectories that correspond to the distinct sub-units of Fig. 8. In these, instead of the different directions (as seen in Fig. 7(c)) we have created sub-units that explain at the same time the direction for each one of the different scales.

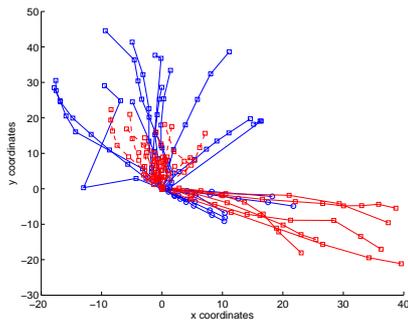


Figure 8: The trajectories for four different sub-units mapped on the 2D signing space represented with different color/marker. Sub-units account for the multiple-cues of both direction and scale of the dynamic segments.

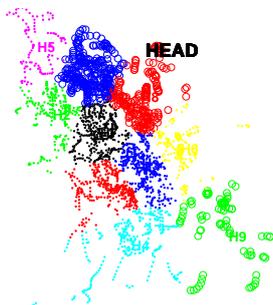


Figure 9: Partitioning of the 2D signing space by K-means. Different colors discriminate the sub-unit.

3.5. Modeling Static Segments

Given the discrimination and separate modeling of the dynamic segments, the remaining segments correspond to the low velocity profiles. We model *only* these static positions and not all positions as they lay across time within movement segments. After applying k-means clustering to the non-normalized positions we get a partitioning of the static positions relative to the head of the signer. Figure 9 shows the constructed sub-units as they are mapped on the 2D signing space.

4. Lexicon: Recombining the Dynamic and Static Segments

4.1. Lexicon Construction

After decomposing the dynamic and static segments for separate modeling, we re-compose them via the lexicon so as to form the complete signs via a concatenation of the sub-units at a symbolic level. Each sub-unit is in this case a ‘symbol’ that is uniquely identified by the feature that has been employed for its construction and the index that has been assigned during the clustering procedure. This lexicon is completely data-driven and does not employ any linguistic information. The lexicon re-composes the two levels of 1) the Dynamic Movement Segments (D) and the 2) Static Position Segments (S). An example of three different lexicons that have been obtained using Position (P) for the static segments and Direction (D) or Movement Trajectories (SPn) or Scale (S) for the dynamic segments respectively is shown in Fig. 10.

BECAUSEP1:HP10	BECAUSEP1:HP10
BECAUSEP2:HP1 MSPn1	BECAUSEP2:HP1 MS1
BECAUSEP3:HP10 MSPn1 HP10	BECAUSEP3:HP10 MS1 HP10
BETTERP3:MSPn1 HP2	BETTERP3:MS1 HP2
BETTERP4:MSPn1 HP8	BETTERP4:MS1 HP8
BIGP1:HP6 MSPn1 HP2	BIGP1:HP6 MS2 HP2
BIGP2:HP8 MSPn1 HP7	BIGP2:HP8 MS1 HP7

Figure 10: Lexicon sample for two different type of features (from left to right) SPn, S for the dynamic and P for static segments.

4.2. Multiple Pronunciations

The realization of signs during continuous natural signing introduces factors that increase the articulation variability. Among the reasons responsible for these multiple pronunciations is the existence of features that do not remain constant during each gloss articulation. For instance there might be cases of the same gloss that is represented by the same sequence of movements but in multiple realizations that involve different initial positions. An example of varying pronunciation for a specific gloss is illustrated by the sample lexicons shown in Fig. 10. Each line in a lexicon sample consists of 1) a gloss identifier concatenated by 2) an index that corresponds to the pronunciation realization case. Figure 10 includes two cases of features for the Dynamic segments combined in all cases with the Position feature for the Static segments. In the shown example, gloss “BECAUSE” is being mapped to three different sub-unit sequences. These specific sub-unit sequences share the first sub-unit of static modeling, while the second one adds a movement sub-unit, e.g. MSPn1, and the third one adds another static sub-unit.

4.3. Sub-Unit sequences to Multiple Glosses Mapping

Among the reasons responsible for these multiple pronunciations is the non-sufficient during this stage of modeling w.r.t. the features employed. For instance there might be cases of glosses that are represented by the same sequence of movements-positions but they involve different handshape configurations that are not accounted yet. Such a case are signs “WITH” and “FOOTBALL” which share common sequence of movements-positions but different handshape configuration. Another factor is the model order we employ, or in other words how loose is the sub-unit clustering we apply. For example if we use a small number of clusters in order to represent all space of movements, although we might have used sufficient features, multiple different movements shall be mapped to the same sub-unit creating looser models.

5. Recognition Experiments

Experimental configuration

In the experiments described we use only the front camera video stream. Among the whole corpus, we restrict our processing on six videos that contain stories narrated from a single signer². We utilize 50 randomly selected signs

²Videos are identified namely as: accident,

among the most frequent ones. We employ 60% of the data for training and 40% for testing. This partitioning samples data from all videos, and among all realizations per sign in order to equalize gloss occurrence. For the evaluation of the recognition results we employ the standard measure of accuracy in the sub-unit level and the gloss level.

Experiments: Next we describe recognition experiments that evaluate the main aspects discussed. 1) We examine the incorporation of the segmentation and classification component referred to as Static vs. Dynamic Classification; this step affects also the adapted modeling w.r.t the employed multiple cues and clustering. 2) Another contribution discussed is the feature normalization for the Dynamic parts that on its turn affects both the modeling and the recognition results. 3) Finally, we further evaluate the incorporation of multiple cues in the Dynamic parts modeling. The employed cues are encoded as Direction (D), Movement Trajectory after scale and initial-position normalization (SPn), Scale (S) and non-normalized Position (P). The results contain both gloss-level and sub-unit level accuracies.

Number of Sub-Units: The number of sub-units we use in each case is depended on the existing experimental dataset and on prior linguistic information. The dynamic segments employ 24 sub-units given motivation on the different type of movements (8 for each of straight or curved or other more complex movements). We use four sub-units for scale modeling and 22 sub-units for the static segments' sub-units which corresponds to different but arbitrary places of articulation. These numbers imply the total number of sub-units employed on each recognition experiment described next and are shown on Table 1. Note that for tasks that are to be compared we employ equal number of sub-units. More sub-units imply a more complex task. Another point to stress, (see also the discussion in Section 4.), is that the gloss level results should be viewed given the "single sub-unit sequence mapping to multiple glosses" due to the missing cues (e.g. handshape). The above gloss accuracy considers a gloss as correct if it exists in the set of targets of the specific sub-unit sequence. This is the case *even* if other glosses are present in the same set. That is, the recognition performance evaluation functions towards our favor even if there are multiple glosses mapped from a specific sub-unit sequence.

Single-Stream Continuous SL Recognition: Here, we evaluate the efficacy of the various movement-position cues employed in single stream recognition experiments and at the same time without incorporating the Dynamic-Static Classification. Figure 11(c,d) shows the results for the four single cue cases: P, D, S and SPn. These results should be seen under the following point of view. The sub-unit accuracy is dependent each time on the complexity of the task: For the case of S the employed number of sub-units is much lower compared to the other single cue cases thus the high performance is for a much easier task (see Table 1).

Dynamic-Static Segmentation and Classification: In this case we compare two variants. The first variant evaluates the modeling that exploits the Dynamic-Static Classifica-

tion (DSC) obtained during segmentation. The second one, corresponds to the case in which we employ only the segmentation *without* the Dynamic-Static Classification (no-DSC) of the segments. For the first case above (DSC) we employ for the Dynamic segments the cues of D, SPn and S. On the contrary for the static segments we employ only the P cue. For the second case of no-DSC all segments share the same cue. For this case among all multiple-cue combinations we show the one that performs best (SPn-S-P). The incorporation of the DSC is encoded in the Fig. 11 by the "+" symbol, e.g. A+B shall correspond to the A cue for the dynamic modeling and the B cue for the static. Where two cues are concatenated by "-" as in A-B, this corresponds to the plain concatenation via multiple streams.

First, we should note that by comparing the single cue experiments with the DSC multiple cue case the latter show improved performance. The overall recognition performance for the DSC case Fig. 11(a,b), outperforms the no-DSC case Fig. 11(c,d). More specifically, using the Position (P) cue naively combined with other features (S, SPn, D) implies increased model variance. On the contrary, see Fig.11(a,b), when the cues (SPn, D, S) are modeled plainly in the dynamic parts and the Position cue (P) is only incorporated on the static modeling the results are improved significantly.

Feature Normalization: The importance of normalization is observed for the no-DSC case since the SPn cue outperforms the non-normalized P cue. For the multiple-cue DSC case on which the P is better incorporated, the SPn+P performs much higher than the non-reported accuracy of P+P (i.e. non-normalized cue in the Dynamic modeling resulting on 38% gloss accuracy).

Multiple Cues in Dynamic Modeling: By incorporating multiple cues in the Dynamic modeling as shown in the DSC case, see for instance D-S+P and SPn-S+P compared to S+P, SPn+P, D+P in Fig.11(a,b), there are slight improvements, that should be considered given the number of sub-units reported in Table 1.

6. Conclusions

We propose a modeling structure that incorporates movement-position cues in an unsupervised manner. Each cue is adopted with the appropriate modeling given its functionality during sign language articulation. The modeling is based on the discrimination between Dynamic and Static cases of the movement-position cues, which provides a segmentation and classification of the segments. Secondly, for each type of modeling we incorporated the appropriate cues after normalization. The dynamic sub-units are constructed by clustering at the *model level*. The evaluation of the proposed multiple-cue modeling approach in recognition experiments on the BU400 continuous sign language corpus shows promising results. However, in order to be able to reach more mature conclusions, we shall 1) incorporate phonological and linguistic information, 2) as well as handshape information, that is currently explored via a model based approach and shall be integrated in a common framework.

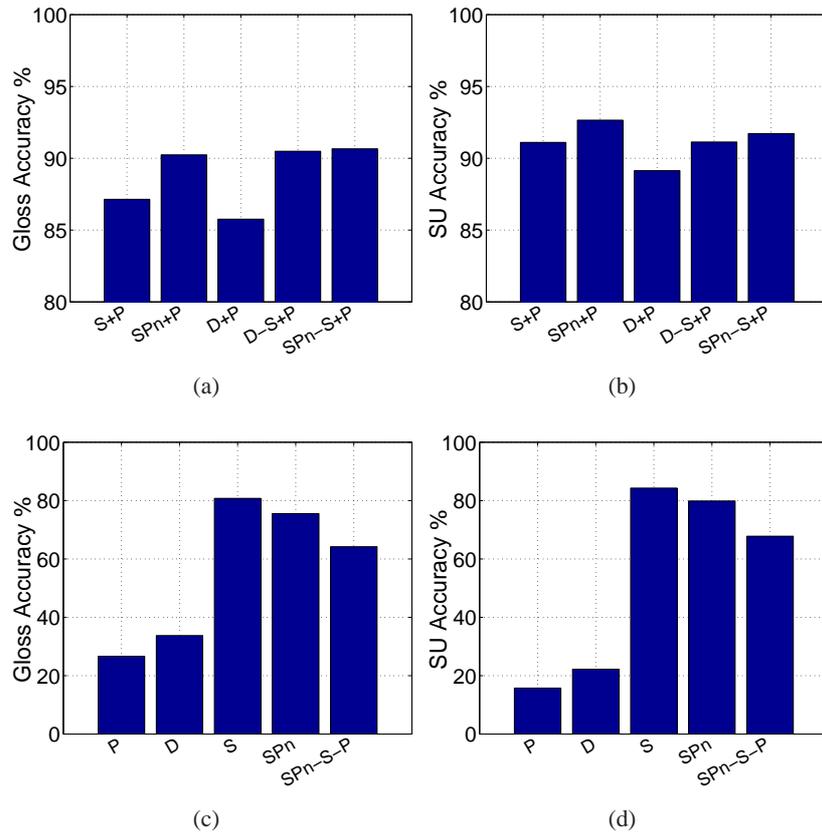


Figure 11: Recognition performance:(a,b) Gloss and Sub-unit accuracy of multiple cues while incorporating Dynamic-Static Classification (DSC), (c,d) Gloss and Sub-unit accuracy of single and one multiple cue without incorporating DSC.

Table 1: Feature identifier corresponding to the recognition experiments and number of sub-units employed.

Feature	S	D	SPn	P	S+P	SPn-S-P	SPn+P	D+P	D-S+P	SPn-S+P
# SUs	4	46	46	46	4+22(46)	24x4+22(118)	24+22(46)	24+22(46)	24x4+22(118)	24x4+22(118)

7. References

- B. Bauer and K. F. Kraiss. 2001. Towards an automatic sign language recognition system using subunits. In *Proc. of Int'l Gesture Workshop*, volume 2298, pages 64–75.
- O. Diamanti and P. Maragos. 2008. Geodesic active regions for segmentation and tracking of human gestures in sign language videos. In *icip*.
- P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *Proc. Int'l Conf. on Language Resources and Evaluation (LREC)*, May.
- K. Emmorey. 2002. *Language, cognition, and the brain: insights from sign language research*. Erlbaum.
- G. Fang, X. Gao, W. Gao, and Y. Chen. 2004. A novel approach to automatically extracting basic units from chinese sign language.
- J. Han, G. Awad, and A. Sutherland. 2009. Modelling and segmenting subunits for sign language recognition based on hand motion analysis. *Pat. Rec. Lett.*, 30(6):623–633.
- B. H. Juang and L. R. Rabiner. 1985. A probabilistic dis-
- tance for hidden markov models. *AT & T Technical Journal*.
- S. K. Liddell and R. E. Johnson. 1989. American sign language: The phonological base. *Sign Language Studies*, 64:195 – 277.
- S. Ong and S. Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *ieeetpami*, 27(6):873–891.
- N. Paragios and R. Deriche. 2002. Geodesic Active Regions: A New Framework to Deal with Frame Partition Problems in Computer Vision. *Journ. of Vis. Commun. and Image Repres.*, 13(1/2):249–268.
- L. R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.
- P. Smyth. 1997. Clustering sequences with hidden markov models. In *In Advances in Neural Information Processing Systems*, volume 9, pages 648–654.
- C. Vogler and D. Metaxas. 2003. Handshapes and movements: Multiple-channel american sign language recognition. In *Gesture Workshop*, pages 247–258.

Sign Language HPSG

Eva Safar, John Glauert

School of Computing Sciences

University of East Anglia

Norwich NR4 7TJ, UK

E-mail: E.Safar@uea.ac.uk, J.Glauert@uea.ac.uk

Abstract

We present an overview of some relevant aspects of sign language synthesis in the ViSiCAST project, which might serve as a possible basis for the Dicta-Sign project. Dicta-Sign is a 3-year EU-funded project, that undertakes parallel corpus collection in different Sign Languages (SLs) and fundamental research and development of sign recognition and generation techniques in order to open up new potential applications for sign language users. One of the aims in Dicta-Sign is to find a model that is suitable for both recognition and generation. In this paper we revisit the main aspects of the synthesis techniques implemented in ALE Prolog using a sign language specific HPSG with the view for future changes needed.

1. Introduction

We present an overview of some relevant aspects of sign language synthesis in the ViSiCAST project¹, which might serve as a possible basis for the Dicta-Sign project. Dicta-Sign² is a 3-year EU-funded project, that undertakes parallel corpus collection in different SLs and fundamental research and development of sign recognition and generation techniques in order to open up new potential applications for sign language users. Therefore the aim in Dicta-Sign is to find a model that is suitable for both recognition and generation.

In the ViSiCAST project we had sound reasons in favour of HPSG (Head-driven Phrase Structure Grammar) for sign language modelling. In sign languages variation in grammars is less than in lexicons, therefore a lexicalist approach is suitable for developing grammars for more than one target languages in parallel. Differences are encoded in the lexicon, while grammar rules are usually shared with occasional variation in semantic principles. A further consideration in favouring HPSG is that the feature structures can incorporate modality-specific aspects (non-manual features) of signs appropriately (Safar & Marshall, 2002).

2. Modifications to the ALE Implementation

Our HPSG is implemented in ALE Prolog (Shieber et al., 1989). The current ViSiCAST feature structure and grammar rules will have to be adapted in a way that they are suitable for analysis as well. Most changes can be

¹ It was an EU Framework V supported project, which developed virtual signing technology in order to provide information access and services to Deaf people.

² We acknowledge that this work is funded through the Dicta-Sign project under the European Union's 7th Framework Programme (grant 231135).

expected in the phonetic and syntactic features. A list of thousand concepts has been collected for parallel corpora in German, French, British and Greek SLs, which initially serve the purpose of guiding the annotation of the collected corpus. This initial lexicon has to be refined to be used in SL grammars enhanced with the linguistic knowledge gained from the corpus for analysis and synthesis.

ALE has been modified to make it compatible with the more recent version of Prolog (SWI v.5.6) on PC and Mac. Picture 1 shows a typical lexical entry for a noun, which will be explained in more detail in the following sections. The left hand side (LHS) represents another modification to ALE. The LHS is a list of HamNoSys transcription symbols for manual and non-manuals (Prillwitz et al., 1989) instead of a word. On the right hand side (RHS) the values of the phonetic (PHON) features are instantiated and propagated to the LHS (like accompanying 'Brow' in this example) via unification and principles. This way we created a dynamic lexicon without increasing compilation time.

3. Architecture in ViSiCAST

ViSiCAST produced a prototype English text to SL translation system. First the English written text was parsed. The output of the parser was then processed using λ -calculus, β -reduction and DRS merging (Blackburn & Bos, 1999). The result was a Discourse Representation Structure (DRS), which in a flattened form served as the input for the HPSG synthesis. In Dicta-Sign after reviving the old system (see Section 2) we can produce the HPSG output. The generated sequence is HamNoSys for manual features and codes for non-manual features. This linguistic analysis can be then linked with the animation technology by encoding the result in XML as SiGML which is then sent to the JASigning animation system (Elliott et al., 2010).

```

[[mug], [Brow], ['mVg', hamfist, hamthumbacrossmod, hamextfingerol, hampalmi,
hamshoulders, hamclose, hamparbegin, hammoveu, hamarcu, hamsmallmod, hamreplace,
hamextfingerul, hampalmdl, hamparend]] --->
word,
  gloss:mug,
  phon: (face:brow:Brow,
        mouth: pict: 'mVg',
        man: (ndh:hns_string,
              hsh: [hamfist, hamthumbacrossmod],
              oxi: (efd: [hamextfingeroll,
                        plm: [hampalmi]),
              mox: [(repeat:R,
                    mloc:Height,
                    m_hns: [hamparbegin, hammoveu, hamarcu, hamreplace,
                           hamextfingerul, hampalmdl, hamparend])],
              const: []),
        allow_weak_drop: minus),
  syn: (precomps: [],
        postcomps: [],
        head: (fix_or_noun, context: (context_in : Çin,
                                     context_out: Çin),
              agr: (cl: (cl_manip: (@ upright_cylinder),
                                   cl_meas: (@ round_obj_meas),
                                   cl_subst: (@ upright_cylinder)),
                    num: (number:Sg, collordist:Coll),
                    per: per_three),
                  form: normal),
              arg_st: [],
              allow_pl_sweep: no,
              allow_pl_repeat: no),
        sem: (index: (ref, Ind),
              count: (number:Sg, collordist:Coll),
              restx: [(sit: Ind3,
                      xeln: mug,
                      sense: 1,
                      args: [])]).

```

Picture 1: An example entry for a noun

4. HPSG Structure

The HPSG feature structure (see Picture 1) starts with the standard PHON (phonetic), SYN (syntactic) and SEM (semantic) components (Pollard & Sag, 1994).

The PHON component describes how the signs are formed by handshape, orientation, finger direction and movement. From the non-manuals the eye-brow movement and mouth-picture were implemented (PHON:FACE:BROW and PHON:MOUTH:PICT).

The SYN component determines the argument structure and conditions for unification. It contains information on what classifiers the word can take (the classifier features are associated with the complements (SYN:HEAD:AGR) and their values are propagated to the PHON structure of the verb in the unification process) or how pluralisation can be realised but also on mode, which is associated with sentence type and (pro)noun drop. The context feature is used to locate things in the three-dimensional signing space. The positions are used for referencing and for directional verbs, where such positions are obligatory morphemes. This feature is propagated through derivation. Movement of objects in signing space is achieved by associating an ADD_LIST and a DELETE_LIST with directional verbs (Safar & Marshall, 2002). Picture 2 shows an example of the HEAD feature of a verb.

The SEM structure includes semantic roles with WordNet definitions for sense to avoid eventual ambiguity in the English gloss.

5. Rules and Principles

The rules deal with sign order of (pre-/post-)modifiers (adjuncts) and (pre-/post-)complements. British Sign Language is a topic-comment language, where the complements can subcategorize for their own complements. Therefore we introduced a Last-Complement rule to finish the recursion of the pre- and postcomp rules. This means that we deviate from the standard Subject-Head rule or schema.

5.1 Mode

The principle of MODE propagates the eye-brow movement's value (neutral, furrowed, raised), which is associated with the sentence type in the input (declarative, yes-no question and wh-question) throughout.

5.2 Prodrop

The second type of principle deals with prodrop, which means the non-overt realization of the pronomina. We introduced an empty lexical entry. The principle checks the semantic head for the values of subject and object prodrop features. Picture 2 shows that the values can be *can* or *can't*, a third value is possible, which is *must*. We then extract the syntactic information for the empty lexical item, which has to be unified with the complement information of the verb. If the value is *can't* prodrop is not possible, in case of *can* we generate both solutions.

```

head:(dirmanipverb_lxm,
  agr:(num:(number:SSg, collordist:Coll),
    per:per, gref:Index2),
  aux:minus,
  context:(context_in : list_context,
    add_list : [ (glossref:[(ref:Index2, glossr:Gloss)],
      locat:locatefd:Efd,
      distance:Distobj,
      heights:Heightobj),
      (glossref:[(ref:Index1, glossr:Glosssubj)],
      locat:locatefd:Efdsubj,
      distance:Distsubj,
      heights:Heightsubj),
      (glossref:[(ref:Index2, glossr:Gloss)],
      locat:locatefd:Efdsubj,
      distance:Distsubj,
      heights:Heightsubj)],
    context_out: list_context,
    delete_list:[(glossref:[(ref:Index2, glossr:Gloss)],
      locat:locatefd:Efd,
      distance:Distobj,
      heights:Heightobj)],
  prodrp_subj:(first:can,
    second:can,
    third:can),
  prodrp_obj:(first:can,
    second:can,
    third:can), %no info on this ~put
  arg_st:[(sem:(index:Index2, Precomp1)), (sem:(index:Index1, Precomp2))]),

```

Picture 2: The HEAD information of a verb

5.3 Plurals

The third type of principle controls the generation of plurals. We handle repeatable nouns and non-repeatable nouns with external quantifiers and plural verbs. The input contains the semantic information that is needed to generate plurals that is a result of the analysis of an English sentence. SLs sign distributive and collective meanings of plurals differently, so the semantic input has to carry that information. English in this respect is often underspecified, therefore in some cases we needed human intervention in the analysis stage. The lexical item determines whether it allows repetition or sweeping movement. Picture 1 shows the `allow_pl_repeat` and the `allow_pl_sweep` feature under SYN. (sweeping movement indicates the collective involvement of a whole group, while repetition has a distributive meaning) When the feature's value is *yes* in any case, then the MOV (movement) feature in PHON is instantiated to the appropriate HamNoSys symbol expressing repetition or sweep motion in agreement with the SEM:COUNT:COLLORDIST feature value. The verb pluralization is handled similarly. For more on plurals, its issues and relation to signing space we refer to (Marshall & Safar, 2005).

5.4 Signing Space

The fourth type of principle is the managing of signing space. In signing, being a visual language, we place objects in the 3D signing space. We not just place them in a certain position but we also move them around. These objects can be referred to via pointing or being signed in the same location (anaphoric relationships), but they can also be manipulated by directional verbs. Directional verbs need a starting and an end point for the movement, which can be obtained by propagating a map of sign space

positions through derivation. The missing phonemes of those positions are available in the SYN:HEAD:CONTEXT feature. While generating the verbs arguments they are populated in different positions of the signing space. If the verb requires the movement of those objects, they will be deleted from the 'old' position and added to a new position. Picture 2 shows the CONTEXT feature with an `add_list` and a `delete_list`. These lists control the changes of the map. The CONTEXT_IN and CONTEXT_OUT features are the initial input and the changed output lists of the map. The map is threaded through the generation. The final CONTEXT_OUT will be the input for the next sentence.

6. How Parameterization works

We will show an example for a lexical entry that has uninstantiated values on the RHS in the PHON structure and therefore the LHS HamNoSys needs to be parameterized as well. (For more details see Marshall & Safar, 2004 and Marshall & Safar, 2005).

```

[[take],
 [Brow],
 ['teIk', Hsh, Efd, Plm, Heightobj, Distobj,
 R1, hamreplace, Efd, Plm, Heightsubj,
 Distsubj, R2]] ---> RHS

```

This is a less frequent example entry when the LHS contains only the HamNoSys structure. The handshape (Hsh), the extended finger direction (Efd) and the palm orientation (Plm) are resolved when the object complement is processed. As in Picture 1 the noun's SYN:HEAD:AGR:CL feature contains information on the different classifier possibilities associated with that noun (@ stands for macro below). In the unification process this information is available for the verb and therefore its PHON features can be instantiated and

propagated to the LHS:

```
syn:(precomps:
  [(@manip(Ph, Gloss, Index2, Precomp1, Hsh,
    Efd, Plm, Sg)),
  (@np2(W, Glosssubj, Plm2, EfdT, Index1,
    Precomp2, Num, PLdistr))]
```

The complements are added to the allocation map (signing space). The allocation map is available for the verb as well which governs the allocation and deletion of places in the map (see SYN:HEAD:CONTEXT feature in Picture 2), therefore the locations for the start and end position can be instantiated in PHON and propagated to the LHS. Heightobj and Distobj stand for the location and the distance from that location for the starting point of the sign, which is the location for the object. Heightsubj and Distsubj stand for the end point of the movement, which is the location of the subject in signing space.

The Brow value is associated with the sentence type in the input and is propagated throughout.

R1 is the placeholder for the sweeping motion of the plural collective reading. R2 stands for the repetition of the movement for a distribute meaning. The verb's SYN:HEAD:AGR:NUM:COLLORDIST feature is unified with the SEM:COUNT feature values. If the SYN:ALLOW_PL_SWEEP or the SYN:ALLOW_PL_REPEAT features permit R1 or R2 can be instantiated according to the semantics. If the semantic input contains singular, R1 and R2 remain uninstantiated and are ignored in the SiGML translation.

7. Conclusion

This approach, i.e. the synthesis within an HPSG framework in a style that allowed to appropriately parameterize the HamNoSys descriptions by inheriting information from other linguistic constructs, proved to be fruitful and could be further developed in the framework of Dicta-Sign.

The Dicta-Sign project undertakes parallel corpus collection and annotation in different SLs and fundamental research and development in a range of (sign recognition and generation) techniques. The lexicon and grammar design therefore have to provide formal representations for recognition, generation and annotation. A lexicon should code information dealing with phonology, semantics, grammar, usage, variation and translation equivalents (compare Johnston,1998). Our HPSG lexicon model in ViSiCAST described signs for intended production providing finer grained details of phonetics and grammar to be able to drive an avatar rather than details of semantics, variation or usage.

The aim in Dicta-Sign is to find a model that is suitable for both recognition and generation. Therefore we have to avoid any specification in the entries, which would restrict recognition, but be specific enough to guide the production. The ViSiCAST grammar was specifically constructed for sign synthesis, so ways to make this process reversible still have to be developed. Also the annotation or translation purposes in different SLs require

more information on variations and exact semantic descriptions.

A list of thousand concepts has been collected for parallel corpora in German, French, British and Greek SLs, which initially serve the purpose of guiding the annotation of the collected corpus. This initial lexicon has to be refined to be used in SL grammars enhanced with the linguistic knowledge gained from the corpus analysis for the purposes explained above.

8. References

- Blackburn, P., Bos, J. (1999). *Representation and Inference for Natural Language. A First Course in Computational Semantics*. Vol II. <http://www.coli.uni-sb.de/~bos/comsem/book1.html>
- Elliott, R., Bueno, J., Kennaway, R., Glauert, J. (2010). Towards the Integration of Synthetic SL Animation with Avatars into Corpus Annotation Tools. In *Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Malta (to be published)
- Johnston, T. (1998). The lexical database of AUSLAN (Australian Sign Language). http://www.sign-lang.uni-hamburg.de/intersign/works_hop1/johnston
- Marshall, I., Safar, E. (2004). Sign Language Generation in an ALE HPSG. In Muller, S. (Ed.), *The Proceedings of the 11th International Conference on Head-Driven Phrase Structure Grammar*. Center for Computational Linguistics (HPSG-2004), Katholieke Universiteit Leuven, pp.189--201, ISSN 1535-1793.
- Marshall, I., Safar, E. (2005). Grammar Development for Sign Language Avatar-Based Synthesis. In *3rd International Conference on UA in HCI*, vol. 8: Universal Access in HCI: Exploring New Dimensions of Diversity, Las Vegas, Nevada, USA, in HCII 2005 (CD-ROM).
- Pollard, C., Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. The University of Chicago Press, Chicago.
- Prillwitz, S., Leven, R., Zienert, H., Hanke, T., Henning, J., et al.(1989). Hamburg Notation System for Sign Languages. An Introductory Guide. *International Studies on sign Language and the Communication of the Deaf*. Vol. 5., Institute of German Sign Language and Communication of the Deaf, University of Hamburg.
- Safar, E., Marshall, I. (2002). Sign Language Translation via DRT and HPSG. In A. Gelbukh (Ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Mexico. Lecture Notes in Computer Science 2276, Springer Verlag, ISBN0302-9743, pp. 58—68.
- Shieber, M., van Noord, G., Moore, C., Pereira, F.C.N.(1989). A Semantic-head-driven Generation Algorithm for Unification-based Formalisms. In *27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, pp. 7—17.

Language Resources for Spanish - Spanish Sign Language (LSE) translation

Rubén San-Segundo¹, Verónica López¹, Raquel Martín¹, David Sánchez², Alfonso García²

¹Grupo de Tecnología del Habla-Universidad Politécnica de Madrid

²Fundación CNSE

Abstract

This paper describes the development of a Spanish-Spanish Sign Language (LSE) translation system. Firstly, it describes the first Spanish-Spanish Sign Language (LSE) parallel corpus focused on two specific domains: the renewal of the Identity Document and Driver's License. This corpus includes more than 4,000 Spanish sentences (in these domains), their LSE translation and a video for each LSE sentence with the sign language representation. This corpus also contains more than 700 sign descriptions in several sign-writing specifications. The translation system developed with this corpus consists of two modules: a Spanish into LSE translation module that is composed of a speech recognizer (for decoding the spoken utterance into a word sequence), a natural language translator (for converting a word sequence into a sequence of signs) and a 3D avatar animation module (for playing back the signs). The second module is a Spanish generator from LSE made up of a visual interface (for specifying a sequence of signs in sign-writing), a language translator (for generating the sequence of words in Spanish) and a text to speech converter. For each language translation, the system uses three technologies: an example-based strategy, a rule-based translation method and a statistical translator.

1. Introduction

In Spain, 92% of the Deaf have a lot of difficulties in understanding and expressing themselves in written Spanish and around 47% of the Deaf, older than 10, do not have basic level studies (information from INE –Spanish Statistics Institute- and MEC –Ministry of Education-). The main problems are related to verb conjugations, gender/number concordances and abstract concepts.

In 2007, Spanish Sign Language was accepted by the Spanish Government as one of the official languages in Spain, and it was defined a plan to invest in new resources in this language. One important problem is that LSE is not disseminated enough among people who can hear. This is why there are communication barriers between deaf and hearing people. These barriers are even more problematic when they appear between a deaf person and a government employee who is providing a personal service, since they can cause the Deaf to have fewer opportunities or rights. This happens, for example, when people want to renew the Identity Document or the Driver's License (DL). Generally, a lot of government employees do not know LSE, so a deaf person needs an interpreter for accessing to these services. Thanks to associations like the Fundación CNSE, LSE is becoming not only the natural language for the Deaf to communicate with, but also a powerful instrument when communicating to people who can hear, or accessing information.

2. State of the Art

The research into sign language has been possible thanks to corpora generated by several groups. Some examples are: a corpus composed of more than 300 hours from 100 speakers in Australian Sign Language (Johnston T., 2008). The RWTH-BOSTON-400 Database that contains 843 sentences with about 400 different signs from 5 speakers in American Sign Language with English annotations (Dreuw et al., 2008a). The British Sign Language Corpus Project tries to create a machine-readable digital corpus of spontaneous and elicited British Sign Language (BSL)

collected from deaf native signers and early learners across the United Kingdom (Schembri, 2008). And a corpus developed at Institute for Language and Speech Processing (ILSP) and that contains parts of free signing narration, as well as a considerable amount of grouped signed phrases and sentence level utterances (Efthimiou E., and Fotinea, E., 2008).

In recent years, several groups have developed prototypes for translating Spoken language into Sign Languages: example-based (Morrissey and Way, 2005), rule-based (San-Segundo et al 2008), full sentence (Cox et al, 2002) or statistical approaches (Bungeroth and Ney, 2004; Morrissey et al, 2007; SiSi system) approaches. About speech generation from sign language, in the Computer Science department of the RWTH, Aachen University, P. Dreuw supervised by H. Ney is making a significant effort into recognizing continuous sign language from video processing (Dreuw et al, 2008b; Dreuw, 2009). The results obtained are very promising.

This paper describes the parallel corpus obtained for developing a Spanish-Spanish Sign Language (LSE) translation system in two specific application domains: the renewal of the Identity Document and Driver's License.

3. Spanish-LSE parallel corpus

The corpus developed in this project has been obtained with the collaboration of Local Government Offices where the mentioned services (the renewal of the Identity Document (ID) and Driver's License (DL)) are provided. The most frequent explanations (from government employees) and the most frequent questions (from the user) were taken down over a period of three weeks and more than 5,000 sentences were noted and analysed.

Not all the sentences refer to ID or DL renewal (Government Offices provide more services), so sentences had to be selected manually. Finally, 1360 sentences were collected: 1,023 pronounced by government employees and 337 by users. These sentences were translated into LSE, both in text (sequence of glosses) and in video, and

50%. For the non-manual part of the sign, the design was always made from scratch, using the tools provided in the Visual Editor.

4. Spanish into LSE translation

The Spanish into LSE translation module is composed of three modules (Figure 3).

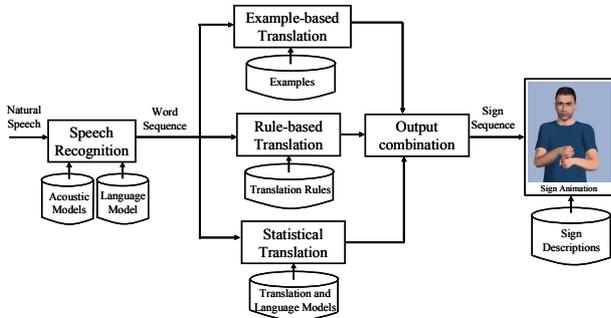


Figure 3: Spanish into LSE translation module

The first one is a speech recognition module that converts natural speech into a sequence of words (text). The second one is a natural language translation module that converts a word sequence into a sign sequence. For each translation, three different strategies are combined at the output step. The first one consists of an example-based strategy: the translation process is carried out based on the similarity between the sentence to be translated and the items of a parallel corpus with translated examples. Secondly, a rule-based translation strategy, where a set of translation rules (defined by an expert) guides the translation process. The last one is based on a statistical translation approach where parallel corpora are used for training language and translation models. We have considered two statistical alternatives: phrase-based one and Finite State Transducers (FST). Table 2, summarizes the results for rule-based and statistical approaches in laboratory tests: SER (Sign Error Rate), PER (Position Independent SER) and BLEU (BiLingual Evaluation Understudy).

		SER	PER	BLEU
Statistical approach	Phrase-based	39.01	37.05	0.5612
	FST-based	34.46	33.29	0.6433
Rule-based approach		21.45	17.24	0.6823

Table 1. Result summary for rule-based and statistical approaches

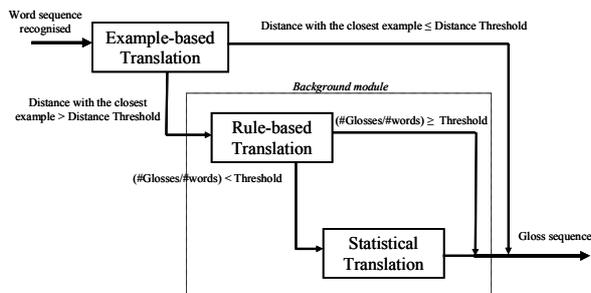


Figure 4: Diagram of natural language translation module combining three different translation strategies

The translation module has a hierarchical structure (Figure 4). Firstly, an example-based strategy is used to translate

the word sequence. If the distance with the closest example is lower than a threshold, the translation output is the same than the example. But if the distance is higher, a background module translates the word sequence, using a combination of rule-based and statistical translators. The last module represents the signs with VGuido (the eSIGN 3D avatar). It is important to remark that this system translate Spanish into LSE, not into Signed Spanish.

5. Spanish generator from LSE

The spoken Spanish generation system is composed of three modules (Figure 5).

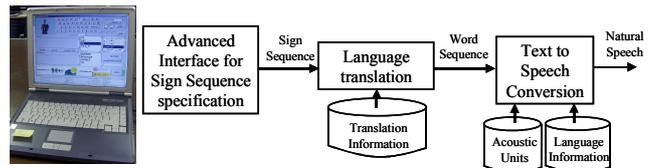


Figure 5: Diagram of Spanish generation system

The first module is an interface for specifying a sign sequence. This interface includes several tools for sign specification: avatar for sign representation (to verify that sign corresponds to the gloss), prediction mechanisms, calendar and clock for date or time definitions, etc. With this visual interface the Deaf can build a sign sentence that will be translated into Spanish and spoken to a hearing person. The sign sequence is specified in glosses but signs can be searched by using specific sign characteristics in HamNoSys notation. (Figure 6)

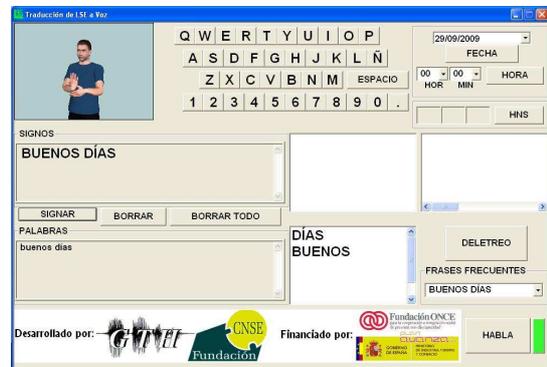


Figure 6: Visual interface for sign sequence specification

The second module converts a sign sequence into a word sequence with three different strategies combined: an example-based, a rule-based and a statistical translation strategy. The procedure is the same as in the Spanish into LSE translation system. The last module converts the word sequence into spoken Spanish by using a commercial Text to Speech converter. In this project the Loquendo system has been used (<http://www.loquendo.com/en/>).

6. Evaluation

An evaluation has been performed for testing the speech into LSE translator and the spoken Spanish generator for Driver's License renewal. The speech-LSE system translates the government employee's explanations and the spoken Spanish generator helps Deaf to ask questions.

The evaluation was carried out over two days. On the first day, a one-hour talk, about the project and the evaluation, was given to government employees (2 people) and users (10 people) involved in the evaluation. Six different scenarios were defined in order to specify real situations. The sequence of scenarios was randomly selected for each user. Ten deaf users interacted with two government employees at the Toledo Traffic Office using the developed system. These ten users (six males and four females) tested the system in almost all the scenarios described previously: 48 dialogues were recorded (12 dialogues were missing because several users had to leave the evaluation session before finishing all the scenarios). The user ages ranged between 22 and 55 years (40.9 average). For both systems the translation accuracy was very high (> 90%) but the users reported several problems related to avatar naturalness and LSE normalization.

7. Conclusion

This paper has described the first Spanish-LSE parallel corpus for language processing research focusing on specific domains: the renewal of the Identity Document and Driver's License. This corpus includes 4,080 Spanish sentences translated into LSE. This corpus also contains a sign database including all sign descriptions in several sign-writing specifications: Glosses, HamNoSys and SEA: Sistema de Escritura Alfabética. This sign database includes all signs in the parallel corpus and signs for all the letters (necessary for word spelling), numbers from 0 to 100, numbers for time specification, months and week days. The sign database has been generated using a new version of the eSign Editor.

This paper also has described the design and development of a Spanish into Spanish Sign Language (LSE: Lengua de Signos Española) translation system. This system is made up of a speech recognizer (for decoding the spoken utterance into a word sequence), a natural language translator (for converting a word sequence into a sequence of signs belonging to the sign language), and a 3D avatar animation module (for playing back the signs). For the natural language translator, three technological proposals have been evaluated and combined in a hierarchical structure: an example-based strategy, a rule-based translation method and a statistical translator.

Finally, this paper has presented a spoken Spanish generator from sign-writing of Spanish Sign Language (LSE: Lengua de Signos Española). This system consists of an advanced visual interface where a deaf person can specify a sequence of signs in sign-language, a language translator (for generating the sequence of words in Spanish), and finally, a text to speech converter.

8. Acknowledgements

The authors want to thank the eSIGN (Essential Sign Language Information on Government Networks) consortium for permitting the use of the eSIGN Editor and

the 3D avatar in this research work. This work has been supported by Plan Avanza Exp N^o: PAV-070000-2007-567, ROBONAUTA (MEC ref: DPI2007-66846-c02-02) and SD-TEAM (MEC ref: TIN2008-06856-C05-03) projects.

9. References

- Bungeroth J., Ney, H.: Statistical Sign Language Translation. In Workshop on Representation and Processing of Sign Languages, LREC 2004, 105-108.
- Cox, S.J., Lincoln M., Tryggvason J., Nakisa M., Wells M., Mand Tutt, and Abbott, S., 2002 "TESSA, a system to aid communication with deaf people". In ASSETS 2002, pages 205-212, Edinburgh, Scotland, 2002
- Dreuw P., Neidle C., Athitsos V., Sclaroff S., and Ney H. 2008a. "Benchmark Databases for Video-Based Automatic Sign Language Recognition". In International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, May 2008.
- Dreuw, P., D. Stein, T. Deselaers, D. Rybach, M. Zahedi, J. Bungeroth, and H. Ney. 2008b "Spoken Language Processing Techniques for Sign Language Recognition and Translation. Journal Technology and Dissability. Volume 20 Pages 121-133. ISSN 1055-4181.
- Dreuw, P., Stein D., and Ney H. 2009. "Enhancing a Sign Language Translation System with Vision-Based Features". LNAI, number 5085, pages 108-113, Lisbon, Portugal, January 2009.
- Efthimiou E., and Fotinea, E., 2008 "GSLC: Creation and Annotation of a Greek Sign Language Corpus for HCI" LREC 2008.
- Johnston T., 2008. "Corpus linguistics and signed languages: no lemmata, no corpus". 3rd Workshop on the Representation and Processing of Sign Languages, June 1. 2008.
- Herrero, Ángel. 2004 "Escritura alfabética de la Lengua de Signos Española" Universidad de Alicante.
- Morrissey S., and Way A., 2005. "An example-based approach to translating sign language". In Workshop Example-Based Machine Translation (MT X-05), pages 109-116, Phuket, Thailand, September.
- Morrissey S., Way A., Stein D., Bungeroth J., and Ney H., 2007 "Towards a Hybrid Data-Driven MT System for Sign Languages. Machine Translation Summit (MT Summit)", pages 329-335, Copenhagen, Denmark, September 2007.
- Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning, et-al. 1989. "Hamburg Notation System for Sign Languages - An introductory Guide". International Studies on Sign Language and the Communication of the Deaf, Volume 5. University of Hamburg, 1989.
- San-Segundo R., Barra R., Córdoba R., D'Haro L.F., Fernández F., Ferreiros J., Lucas J.M., Macías-Guarasa J., Montero J.M., Pardo J.M., 2008. "Speech to Sign Language translation system for Spanish". Speech Communication, Vol 50. 1009-1020. 2008.
- Schembri, A., 2008 "British Sign Language Corpus Project: Open Access Archives and the Observer's Paradox". Deafness Cognition and Language Research Centre, University College London. LREC 2008.
- Zwitterslood, I., Verlinden, M., Ros, J., van der Schoot, S., 2004. "Synthetic Signing for the Deaf: eSIGN". Workshop on Assistive Technologies for Vision and Hearing Impairment, CVHI 2004, 29 June-2 July 2004, Granada, Spain.

Issues in creating annotation standards for sign language description

Adam Schembri¹, Onno Crasborn²

¹Deafness Cognition and Language Research Centre, University College London
49 Gordon Square, London, WC1H 0PD, United Kingdom

²Centre for Language Studies, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: a.schembri@ucl.ac.uk; o.crasborn@let.ru.nl

Abstract

In this paper, we discuss the need for a standardised system of annotation for sign language corpora. Although several tools exist for the annotation of video data (such as ELAN or iLex), and some existing projects have annotation guidelines (e.g., Crasborn et al., 2007; Johnston, 2010), a widely adopted standard is currently unavailable. First, we discuss the purpose of a set of unified annotation standards for sign languages: such standards would provide a shared set of conventions for the easy exchange of data across different sign language corpus projects and may increase consistency within corpora. Next, we discuss the properties that would define a good set of shared annotation conventions (Beckman et al., 2009). We examine some of the proposed annotation standards for spoken language description, such as the ToBI conventions for prosody and the Leipzig Glossing Rules for morpho-syntax. Lastly, we discuss the relationship between theory and description. Dryer (2006) pointed out that linguists often contrast ‘theoretical linguistics’ with ‘descriptive’ work. But if one accepts the argument that there is indeed no ‘atheoretical description’, then sign language linguists need to agree on a shared theory for basic sign language description, and how this translates into annotation practices.

1. Introduction

In this paper, we discuss the need for standardised annotation conventions for the creation of signed language corpora. The paper has come about partly in response to an increasing interest in annotation standards amongst spoken language linguists, as manifested in the report by the annotation standards working group at the 2009 Cyberling workshop (Beckman et al., 2009), as well as among some sign language researchers (e.g., Hermann, 2008; Johnston, 2010). Annotation is used here to refer to written material that is added to, and time-aligned with, the primary sign language digital video data, and represents a description and/or an analysis of the data. Several multimedia annotation tools are currently available (e.g., ELAN, iLex, Anvil, Transana, and SignStream), and are increasingly becoming adopted in the sign language linguistics community. Despite the fact that sign language researchers form a relatively small community of practice and that some projects have made their annotation guidelines available (e.g., Neidle, 2002; Crasborn et al., 2007; Zwitserlood et al., 2008), widely accepted conventions for sign language transcription and annotation are lacking. In the absence of any agreed set of standards, the conventions adopted by the ECHO project¹ have become the basis for some researchers’ annotation guidelines (e.g., Johnston & Schembri, 2006; Herrmann, 2008; Leeson & Nolan 2008), but we feel that the time for wider discussion and dissemination of an

agreed set of standards has come.

Note that we are not proposing the widespread adoption of any sign language writing or notation system, nor for a movement away from the increasing use of primary video data in the field: we are focusing here on the use of annotation as means of tagging the primary data and allowing us to create machine-readable corpora.

2. Sign language annotation

Ide and Romary (2004) suggested that there are two fundamental types of annotation activity: (1) *segmentation* and (2) *linguistic annotation*. The first activity consists of identification of the observable elements in the primary data (e.g., signs) using glosses, and should involve some kind of *tokenisation* or *lemmatisation* of the data (Johnston, 2010). The second activity might be further subdivided into at least two subtypes: *syntagmatic* and *paradigmatic* annotation (Beckman et al., 2009). Syntagmatic annotation involves a description of the relationship between the elements identified in the segmentation process (e.g., a noun phrase), while paradigmatic annotation involves the identification of segments as members of particular linguistic classes (e.g., nouns or verbs). Sign language glossing techniques used in the literature often attempt to combine all of these aspects into a single string (e.g., glosses representing signs combined with class labels, such as ‘CL’ for classifier, and superscript lines showing the scope of non-manual markers, such as ‘neg’ for a headshake over a verb phrase).

3. Why do we need sign language

¹ <http://www.let.ru.nl/sign-lang/echo>

annotation standards?

Annotation of sign language video data serves a number of different functions in corpus sign linguistics, reflecting a researcher's interest in the specific phonetic, phonological, lexical, morphological, syntactic and/or discourse organisation of the data. Often, annotation guidelines are created to serve very specific purposes. In the current British Sign Language (BSL) Corpus Project, for example, a study investigating the linguistic and social factors influencing variation in signs produced with the 1 handshape (the index finger extended from the fist) uses dedicated single character codes for each of the relevant factors, such as the handshape in the preceding sign, or the gender of the signer (Schembri et al., 2009). Annotation conventions will thus always be complemented by project-specific annotations, and are by no means intended to replace these.

The issue of annotation standards becomes more important as opportunities for researchers to share data grow. As Johnston and Schembri pointed out (in press), very few sign language corpora in the modern sense of the term 'linguistic corpus' currently exist (i.e., a representative collection of language samples in a machine-readable form that can be used to study the type and frequency of linguistic units, see McEnery & Wilson, 2001). But many corpus projects are now underway, and this provides the field with a window of opportunity to address the issue of annotation standards. We should begin focussing on the issue of standardised conventions now to ensure that future data exchange between these various projects will be possible, and to provide a basis for future projects. Beckman et al. (2009) suggested that an annotation standard will only succeed if it is associated with a commitment by a community of users to adhere to such conventions. As more and more sign language researchers begin to work on similar issues in corpus sign linguistics, meet regularly in specific workshops and share resources through the Sign Language Linguistics Society² and the Sign Linguistics Corpora Network³, there are now structures in place that can support the development, codification and transmission of annotation standards.

Aside from being able to exchange data between corpora, annotation standards might also encourage consistency within corpora. Good standards will be based on experiences from multiple researchers and research areas and are more likely to have well-developed manuals for annotators or other training methods like dedicated workshops.

4. What are the characteristics of best practice annotation standards?

Beckman et al. (2009) proposed a number of properties as features of 'best practice' annotation standards. First, standards have to be *consistent* and *reliable*. If we look at the history of sign language representation practices in

the sign language literature, there have been few attempts to evaluate the reliability of our means of representing sign language data (such as glossing). This is because there have been few opportunities for sharing primary data, and thus issues around the reliability of particular practices have been avoided. Thus, in order to ensure consistency and reliability for any proposed set of standards, there may be a need to conduct studies into the intra-annotator and inter-annotator reliability rates of any such system, and structures in place that will allow revisions of the standards to be disseminated. Independent validation of a whole corpus is impossible if there is not explicit agreement on the annotation standards that should apply and if these standards are not described in detail.

Second, standards should be *useable*. Any proposed set of conventions must be accompanied by extensive documentation (e.g., reference and training manuals) and perhaps specially-designed annotation software, be relatively easy to teach, should allow the data that has been annotated to be searched using already available query tools, and should comply with the technical demands of a specific annotation tool (e.g. on the text encoding standard to follow).

Third, annotation conventions should be *resilient*. Often there may be uncertainty about how best to annotate some aspect of the primary data, so the standards need clear mechanisms for marking uncertainty about ambiguous cases.

Fourth, standards should be *accountable*. The amount of information contained in the annotations, for example, should stay within the limits of confidentiality agreed to by corpus participants.

Fifth, annotation conventions need *interoperability*: the standards need to be useable within different annotation software packages. They must be clearly related to existing descriptions of the specific linguistic phenomena in the literature, and users should be able to translate the annotation conventions into the terminology used by their own particular theoretical framework.

Lastly, the standards need *extensibility* and *adaptability*. The annotations should be able to be extended to describe new linguistic phenomena in undocumented sign language varieties. There also need to be practices related to versioning the conventions, so that metadata about which version of the standards are used in particular corpora are available, together with mechanisms for translating across corpora that have been annotated at different stages during the evolution of the conventions.

5. Case studies of spoken language annotation standards

Beckman et al. (2009) review many of the existing standards for annotation for spoken languages. Two examples that illustrate different aspects of the issues involved in the creation of standardised annotation conventions include the Leipzig Glossing Rules and the ToBI Framework.

² <http://www.slls.eu>

³ <http://www.ru.nl/slcn>

5.1 Leipzig Glossing Rules⁴

The Leipzig Glossing Rules⁵ are a de facto standard for glossing morphosyntactic phenomena proposed by linguists at the Max Planck Institute for Evolutionary Anthropology and the University of Leipzig. The conventions have emerged out of the typological literature, building on work by Lehmann (1983) and Croft (2003). The rules includes recommendations for best practice with interlinear glosses, such as a requirement for word-by-word alignment of glosses with words, with segmentable morpheme glosses separated by hyphens and fused morphemes represented by glosses separated by periods. Infixes are shown using angled brackets in the gloss, and reduplication shown by a tilde. The rules also include a lexicon of abbreviation conventions for various morphosyntactic categories. These include labels such as ‘AGR’ for agreement markers, ‘OBL’ for oblique arguments and ‘VOC’ for vocative constructions. The rules reflect common usage in the typological literature (and indeed some of the practices and labels will be familiar from published sign language research), with only a few innovations proposed.

Documentation consists of a website, with the rules downloadable as a PDF document. Feedback is welcome, with possible revised versions of the rules promised for the future (the current version dates from February 2008), but currently there is little information available about the consistency and reliability of their use. Beckman et al. (2009) suggest that the creation of some software that allowed users to check their annotations for internal consistency would be useful.

5.2 ToBI

Unlike the Leipzig Glossing Rules, the ToBI (Tone and Break Indices) conventions were originally language-specific, intending to work as a set of annotation standards for the description of the prosody and intonation of American English. This has since been extended to other varieties of English and to a number of other spoken languages. Although these different systems share some basic design principles, they are language-specific, as each set of annotation conventions ‘must be guided by an inventory of its prosodic and intonation patterns’ (Pierrehumbert, 2000: 26)⁶.

Nevertheless, the standards to provide a basis for comparing prosodic systems across languages using shared terminology. A ToBI annotation for American English includes six obligatory parts (Beckman et al., 2005): (1) an audio recording, (2) a record of the fundamental frequency contour, (3) an autosegmental transcription of the intonation contour, (4) an representation of each lexical item, (5) a numeric index from 0 to 4 of the perceived degree of juncture after each

lexical item, and (6) markers for disfluencies, commentaries and other miscellaneous annotations. Symbols include L and H for low and high tones, with % representing boundaries, and ? for uncertainty about the annotation. The system for English represents a consensus model of intonation and prosody, drawing on common elements in the 80 years of inter-disciplinary basic and applied research into English prosody. ToBI has had considerable development, testing and a history of use since the early 1990s. Documentation includes websites⁷ and published articles, and there have been a number of workshops held at international conferences.

6 Theory and sign language description: Implications for sign language annotation standards

An issue that has been clearly stated in the work on prosodic systems (Beckman et al., 2005) and morphosyntax (Dryer, 2006) is that a theory-neutral annotation system is impossible. Beckman et al. (2005) pointed out that even the most widely-accepted annotation standard, the International Phonetic Alphabet (IPA), is based on two strong theoretical claims: that utterances in any spoken language can be divided into basic vowel and consonant segments (rather than taking syllables as the basic smallest unit, for example), and that each spoken language has a limited inventory of speech sounds that are not radically different from the languages on which the IPA was initially based. Dryer (2006) pointed out that linguists often characterise certain work as ‘atheoretical’, with some researchers, for example, contrasting ‘theoretical linguistics’ with ‘descriptive’ work on particular languages or in cross-linguistic typology. But if one accepts the argument that there is indeed no ‘atheoretical description’, then sign language linguists need to agree on what sort of shared theory we need for basic sign language description, and how this translates into annotation practices. This will be a challenge, particularly in sign language morphology, where, for example, there is a lack of consensus in the field about whether or not signed languages have verb agreement (e.g., Liddell, 2000; Meier, 2002) and verbal classifier systems (e.g., Schembri, 2003; Zwitserlood, 200x).

7 Towards annotation conventions

We can see the beginnings of standardised annotation conventions for sign language corpora in the ECHO project (Crasborn et al., 2007). The ECHO guidelines were the outcome of a pilot project on the creation of open access sign language corpora on the internet, in which researchers from three universities in different countries and with different research interests aimed to establish a set of basic annotation layers that would be of use for various research endeavors in the future. This led to annotation guidelines and a set of short annotated narratives and poetry (Crasborn et al., 2007). The

⁴ <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

⁵ <http://www.eva.mpg.de/lingua/resources/glossing-rules.php>

⁶ See for example <http://todi.let.kun.nl/ToDI/home.htm> for the Transcription Of Dutch Intonation (TODI).

⁷ See for example <http://www.ling.ohio-state.edu/~tobi/>.

annotation layers (tiers in ELAN) included glossing separately for the left and right hand, some phonetic annotations appended to the glosses, and a selection of articulatorily independent non-manual properties such as broad categories for eye blinks and head movements (Nonhebel et al., 2004a). Separate conventions were created for the annotation of mouth actions (Nonhebel et al., 2004b). These proposals have influenced the work on the Auslan, ISL, NGT and BSL corpus annotation guidelines, as well as those used in more specific cross-linguistic projects (e.g., Herrmann, 2008; Zwitserlood et al., 2008). The ECHO project, and subsequent work by Zwitserlood et al. (2008), for example, proposed that terminology for segmentation and linguistic annotation has to be very general, and these suggestions will serve as a basis for future work. It is not sufficient, however, for single individuals or research groups to propose standardised conventions, as any annotation standards must develop out of some consensus view about what aspects of sign language linguistic theory and description are important.

8. Practical implications for sign language annotation standards

It is clear that the creation of standards will require a substantial effort on the part of the corpus sign linguistics community. The field lacks the long tradition and widespread shared terminology that forms the basis of the Leipzig Glossing Rules for morphosyntax, and has not experienced the widespread movement towards the creation of consensus-based conventions that we see in the ToBI standards. Despite this, current infrastructure in the field would lend itself to the creation and dissemination of any such proposed standards for sign language annotation. Metadata standards for sign language corpus work already exist (Crasborn & Hanke, 2003), for example, and to appear to be gaining acceptance amongst sign language researchers.⁸

There clearly appears to be the need for dedicated funding beyond the current Sign Linguistics Corpora Network to support a project focused on the creation of annotation standards, and the preparation of necessary documentation that can be distributed to potential users. Any annotation-related project would also possibly require studies into intra-annotator and inter-annotator reliability, as well as the creation of computational tools that can increase the reliability of annotators' work. Moreover, the large-scale validation of whole corpora will be dependent on well-documented annotation conventions, and the validation process would be of a higher standard if the annotation can indeed rely on shared standards. Moreover, any such project needs to put into place some kind of institutional framework for the ongoing maintenance of the conventions, to provide training, and to support ongoing revisions of the conventions and of the accompanying documentation.

⁸ This early standard on sign metadata has recently been re-evaluated at a workshop of the Sign Linguistics Corpora Network, see <http://www.ru.nl/slcn>.

Finally, it would be a good idea to explore to which extent the standardisation efforts currently encouraged by the pan-European CLARIN project⁹ could be employed. This especially holds for the standard data categories that define widely agreed-upon linguistic terms in the ISOcat¹⁰ concept registries. These might contribute to conventions for sign language annotation, while at the same time maintaining strong links with the spoken language research domain.

9. Acknowledgements

This work was supported in part by the Economic and Social Research Council of Great Britain (RES-062-23-082), by the Netherlands Organisation for Scientific Research (NWO 236-89-002), and by the European Research Council (ERC Starting Research Grant awarded to Onno Crasborn).

10. References

- Dryer, M.S. (2006). Descriptive theories, explanatory theories, and basic linguistic theory. In F. Ameka, A. Dench & N. Evans (Eds.), *Catching language: Issues in grammar writing*. Berlin: Mouton de Gruyter, pp. 207-234.
- Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005) The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology. The phonology of intonation and phrasing*. Oxford: Oxford University Press, pp. 9-54.
- Beckman, M., Robinson, S., Chung, S., Corbett, G., Fillmore, C., & Wright, R. (2009). *Annotation standards*. Retrieved on 24 March 2010 from the Cyberling Wiki: <http://cyberling.elanguage.net/page/Group+1%3A+Anotation+Standards>.
- Crasborn, O. & Hanke, T. (2003). *Metadata for sign language corpora*. Online document, http://www.let.ru.nl/sign-lang/echo/docs/ECHO_Metadata_SL.pdf
- Crasborn, O., Mesch, J., Waters, D., Nonhebel, A., van der Kooij, E., Woll, B., & Bergman, B. (2007). Sharing sign language data online: Experiences from the ECHO project. *International Journal of Corpus Linguistics*, 12(4), pp. 537-564.
- Croft, W. (2003). *Typology and universals*. 2nd ed. Cambridge: Cambridge University Press.
- Herrmann, A. (2008). Sign language corpora and the problems with ELAN and the ECHO annotation conventions. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 68-73.
- Ide, N. & Romary, L. (2004). International standard for a linguistic annotation framework. *Journal of Natural Language*

⁹ <http://www.clarin.eu>

¹⁰ <http://www.isocat.org>

- Language Engineering*, 10:3-4, 211-225.
- Johnston, T. (2010a). *Guidelines for the annotation of the video data in the Auslan corpus*. Unpublished manuscript, Macquarie University.
- Johnston, T. (2010b). From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1), pp. 106-131.
- Johnston, T. & Schembri, A. (in press). Corpus analysis of sign languages. In C. A. Chapelle, (Ed.), *Encyclopedia of Applied Linguistics*. Wiley-Blackwell.
- Leeson, L. & Nolan, B. (2008). Digital deployment of the Signs of Ireland Corpus in E-learning. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 112-122.
- Lehmann, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica* 16: 199-224.
- Liddell, S.K. (2000). Indicating verbs and pronouns: Pointing away from agreement. In K.D. Emmorey & H. Lane (Eds.), *The Signs of Language revisited: An anthology to honor Ursula Bellugi and Edward Klima*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 303-320.
- McEnery, T. & Wilson, A. (2001). *Corpus Linguistics*, 2nd ed. Edinburgh: Edinburgh University Press.
- Meier, R.P. (2002). The acquisition of verb agreement: pointing out arguments for the linguistic status of agreement in sign languages. In G. Morgan & B. Woll (Eds.), *Directions in sign language acquisition*. Amsterdam, John Benjamins, pp. 115-142.
- Neidle, C. (2002). SignStream™ Annotation: Conventions used for the American Sign Language Linguistic Research Project. [file://localhost/Online report](file://localhost/Online_report). <http://www.bu.edu/asllrp/asllrpr11.pdf>.
- Nonhebel, A., Crasborn, O. & van der Kooij, E. (2004a) Sign language transcription conventions for the ECHO project. Version 9, 20 January 2004. Radboud University Nijmegen. http://www.let.ru.nl/sign-lang/ECHO/docs/ECHO_transcr_conv.pdf.
- Nonhebel, A., Crasborn, O. & van der Kooij, E. (2004b). Sign language transcription conventions for the ECHO project. BSL and NGT mouth annotations. Radboud University Nijmegen. http://www.let.ru.nl/sign-lang/ECHO/docs/ECHO_transcr_mouth.pdf
- Pierrehumbert, J. (2000) Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and experiment: Studies presented to Gosta Bruce*. Dordrecht, Netherlands: Kluwer, pp. 11-26.
- Schembri, A., Fenlon, J., & Rentelis, R. (2009). Sociolinguistic variation in the 1 handshape in British Sign Language. Paper presented at *NWAV 38: The 38th New Ways of Analyzing Variation Conference*, University of Ottawa.
- Schembri, Adam. (2003). Rethinking 'classifiers' in signed languages. In K.D. Emmorey (Ed.), *Perspectives on classifier constructions in sign languages*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 3-34.
- Zwitterlood, I. (2003). *Classifying hand configurations in Nederlandse Gebarentaal (Sign Language of the Netherlands)*. Utrecht, The Netherlands: LOT.
- Zwitterlood, I., A. Özyürek & P. Perniss (2008) Annotation of Sign and Gesture Cross-linguistically. In O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd (Eds.), *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, ELDA, Paris, pp. 185-190.

Synthetic Corpora: A Synergy of Linguistics and Computer Animation

Jerry Schnepf, Rosalee Wolfe, John McDonald

DePaul University

243 S. Wabash Ave, Chicago IL. U.S.A.

E-mail: jschnepf@cdm.depaul.edu, rwolfe@depaul.edu, jmcDonald@cs.depaul.edu

Abstract

Synthetic corpora enable the creation of computer-generated animations depicting sign language and are the complement of corpora containing videotaped exemplars. Any design for a synthetic corpus needs to accommodate linguistic processes as well as support the generation of believable, acceptable synthesized utterances. This paper explores one possibility for representing linguistic and extralinguistic processes that involve the face and reports on the outcomes of a user test evaluating the clarity of utterances synthesized by this approach.

1. Introduction

Synthetic corpora are computer representations of linguistic phenomena. They enable the creation of computer-generated animations depicting sign languages and are the complement of corpora containing videotaped exemplars.

Synthetic corpora have the potential to serve multiple disciplines. They can aid in the automatic recognition of sign (Farhadi, et al., 2007) because they contain the geometric data required for intelligent visual detection algorithms. They can also provide visual depictions of abstract representations and act as a verification tool for data integrity and hypothesis testing (Hanke & Storz, 2008).

Synthesized signs can be modified as they are formed. This provides the flexibility to generate an endless variety of utterances not possible with recordings and opens possibilities for automatic translation efforts. While representing sign for this purpose is still an open question, a synthetic corpus has the potential to serve in this capacity. The flexibility of synthetically-generated sign is also useful for the development of interpreter training software and self-directed learning tools for deaf children (Wolfe, 2006; Wolfe, et al., 2007)

The following describes a design for a synthetic corpus of American Sign Language. In addition to representing glosses, the corpus provides for facial nonmanual signals and extralinguistic facial communication. The paper also reports on a user evaluation of animations generated by this approach.

2. Design Goals

From an animator's perspective, utterances in sign are comprised of geometric poses and movements. Given the proper videotaped reference material, it is possible to animate any signed utterance. However, the animation does not take into account linguistic structure. Whereas the production of computer generated animation only requires timing and geometric data, the synthesis of sign requires additional information, because what is manifested physically is often the result of co-occurring linguistic and extralinguistic processes (Wilbur, 2000).

Figure 1 depicts the gloss BOOK being signed in a yes-no question with happy affect. These co-occurring functions require representation as independent entities so that they can be recombined and thus interact with each other. They have parallels to tracks used in sign annotation software (Brugman & Russell, 2004). Linguistic annotations can help animation transcribers understand the salient features of movements and poses, helping them to build far more legible animations. Thus the classification of geometric changes based on their linguistic function is mandatory for producing novel utterances.

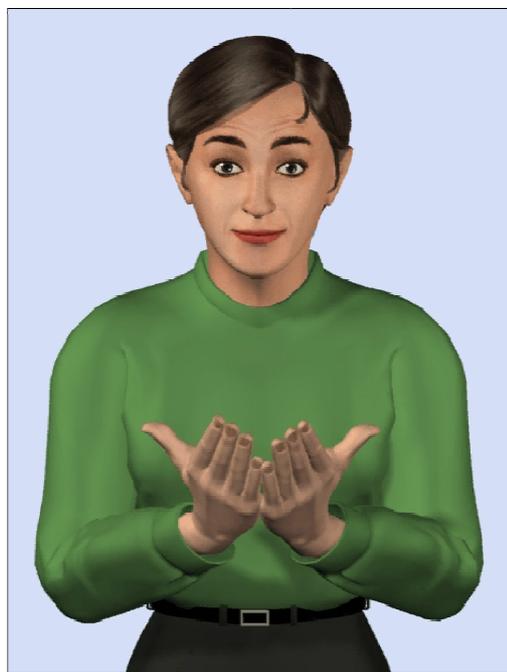


Figure 1: A happy signer asking a Yes/No question.

A desirable feature of any representation is the ability to accommodate paralinguistic and extralinguistic information. Emotional affect must be considered, as well as such phenomena as mouthing, which some populations may prefer. Researchers, however, should have the option to include or exclude this additional data when generating utterances.

To demonstrate the importance of this design goal, consider a Wh-question signed in an angry fashion, as in Figure 2. The eyebrows lower as part of producing a Wh-question. However, the emotional state of anger also involves lowering the eyebrows. The synthesis of this sentence requires that these two be depicted simultaneously.



Figure 2: An angry signer asking a Wh-question.

At first glance, the design goals of linguistics and animation would appear to be at cross purposes. Linguistic researchers often use corpora to form hypotheses through queries on linguistic features, and are interested in such abstractions as phonemes, lexical modifiers and verb agreement. In contrast, animators require extensive minute detail.

In actuality, the fields of linguistics and computer animation create a mutually beneficial synergy. Having the detailed precision required for animation can facilitate the exploration of subtle interactions among linguistic phenomena. Likewise, animators need an abstract representation to organize, combine, and synthesize complex animation data.

Regardless of the animation technique, linguistic knowledge is necessary to produce any synthetic corpus. Animators who hand-transcribe need to work closely with linguists, so that phenomena are tagged correctly. Linguistic information guides the transcription artist's efforts to produce a natural exemplar that encapsulates the essential motions of a sign.

With motion capture, the role of linguistics is no less central. Motion capture equipment generates massive amounts of data that must be cleaned to remove extraneous noise. The linguistic attributes of a sign give the cleanup artists precisely what they need to process and extract the desired motion.

3. Current Proposal

Our work uses labeled manual transcription to create detailed and accurate animations of sign. These animations require voluminous data, as they must be realistic enough to pass the scrutiny of fluent signers. However, such detail is organized using a framework that is both abstract enough to facilitate linguistic research and flexible enough to allow for the synthesis of novel utterances.

Table 1 shows the high level structure of our corpus design, which is based on abstractions used by linguists and is encoded as XML (DuCharme, 1999). High level tracks separately control the linguistic functions of gloss, syntax, and nonmanual lexical modifiers. These direct the position and timing of subordinate geometric components. Researchers have the option to add high level tracks for paralinguistic or extralinguistic functions.

High Level Tracks Linguistic: syntax gloss lexical modifier Extralinguistic: affect mouthing	NM Lexical Modifier Block Label Start time End time Curve Viseme *(multiple possible) Label Time Geometry groups Controllers Keys
Syntax Block Label Start time End time Curve Geometry groups Controllers Keys	Affect Block Label Start time End time Curve Geometry groups Controllers Keys
Gloss Block Label Start time End time Linguistic Component Block Left Handshape Label Time Geometry groups Controllers Keys Right Handshape Label Time Geometry groups Controllers Keys Geometry groups Controllers Keys	Mouthing Block Label Start time End time Curve Viseme *(multiple possible) Label Time Geometry groups Controllers Keys

Table 1: Corpus Structure.

Each track contains blocks of time-based information. Each block has a label, a start time, an end time, as well as a collection of subordinate geometry blocks. Geometry blocks can contain animation keys or a static pose. Further, blocks can contain intensity curves that control the onset and intensity of a pose, allowing for multifarious variations.

Figure 3 demonstrates the abstraction of linguistics and the detail of animation in the case of the question “Do

you want a book?” The green curve represents the movement corresponding to the yes-no question syntactic marker. The red curve represents the influence of the affect “anger”.

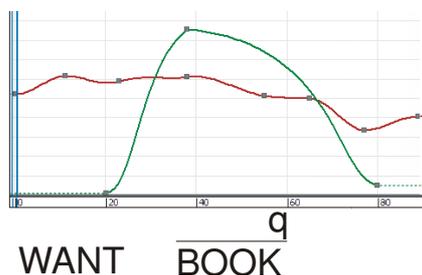


Figure 3: Intensity curves and corresponding sentence.

Although the syntactic marker co-occurs with the gloss BOOK, the green curve controlling the intensity of the corresponding pose starts before the onset of the syntactic marker and ends a significant amount of time after it. This reflects a commonly-used technique in animation whereby the action will ease-in and ease-out of a given pose (Burtnyk & Wein, 1976). Further, animation principles require that the pose not be held perfectly still at any time, thus there is no plateau in the curve.

The use of labeled poses follows common practice in animation studios where a master animator creates a dictionary of characteristic poses (Thomas & Johnston, 1981). By encapsulating minute geometric arrangements in concise groups called poses, a master animator provides an efficient mechanism for others to apply and combine poses. In a similar fashion, this corpus design allows for application and composition of linguistic processes.

4. A Case Study

To test the feasibility of this approach, we focused on the interaction of processes that take place on a signer’s face. We based the design on the substantial body of literature that characterizes these processes (Grossman & Kegl, 2006; Reilly, et al., 1990; Weast, T., 2008). We also considered the feasibility of incorporating both linguistic and extralinguistic information in the design.

We conducted a study of the clarity and acceptability of the synthesized utterances. Since we aimed to represent the interactions of both linguistic and extralinguistic facial movements, we chose a set of test utterances that combined the effects of a single facial linguistic marker and a single emotive pose (See Table 2).

Twenty participants, all of whom were attending the 2009 DeafNation Expo trade show in Palatine, Illinois volunteered to participate in this study. The participants answered background questionnaires to determine their level of ASL fluency. They were informed that they could withdraw at any time during the experiment and they were naive as to its purpose. This work was reviewed and approved by the Institutional Review Board at DePaul University [JS101609CDM].

During the user test, participants viewed animations

of ASL signs. During each session the participant watched short clips depicting the combination of nonmanual signals and emotional affect, as listed in Table 2. The clips are available at <http://asl.depaul.edu/LREC2010>. Following each clip, participants answered questions regarding its meaning and clarity.

\overline{t} BOOKS YOU WANT \overline{WHq} (1) Happy (2) Angry
\overline{CHA} COFFEE TALL (3) Happy (4) Angry

Table 2: Test utterances.

The test environment comprised a PC laptop placed on a table in an exhibition booth. The test facilitator operated the laptop while the participant watched an attached monitor. The participants viewed animations full-screen on the 21” LCD monitor (resolution: 1280 x 1024 pixels). They were seated at a viewing distance of 20-40”. All instructions were signed by the Deaf facilitator or the interpreter. A note-taker sat behind both the participant and facilitator while the interpreter sat across the table.

Each participant tested individually. Participants were informed that they should watch each animation carefully and that they could watch an animation as many times as they wanted.

The facilitator prefaced each animation with a short sentence establishing its context. For example, the first animation displayed “How many books do you want?” Before playing the animation the facilitator explained that the character is the owner of a book store who is taking an order from a customer.

After watching an animation, each participant answered four questions. The first question asked the participant to repeat the sentence to confirm that the animation had communicated the intended meaning. Question two presented a graphical Likert scale (Figure 4) which queried the perceived emotional state. The third question employed another Likert scale measuring the animation’s clarity, from unrecognizable (1) to perfectly clear (5). The last question asked for suggestions to improve the animation.

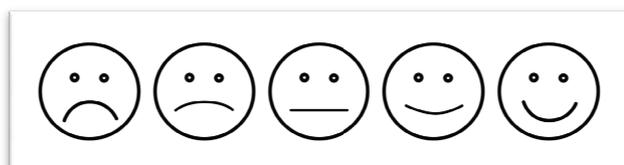


Figure 4: Likert scale measuring emotional state.

5. Results

For brevity, only responses to utterance (4) are reported here. All the results were similar and the entire data set is available at <http://asl.depaul.edu/LREC2010>. In

response to the first question, participants were able to replicate the utterance 100% of the time. Also, 70% rated the animation as clear or very clear (Table 3). Each participant ascertained that the mouth shapes which characterize CHA indicate a large size. While some were confused as to the reason why the avatar appeared angry about a large cup of coffee, 95% correctly identified the intended emotional state (Table 4). After viewing the animation, participants described her as “grumpy”, “angry”, “disappointed” and “negative”.

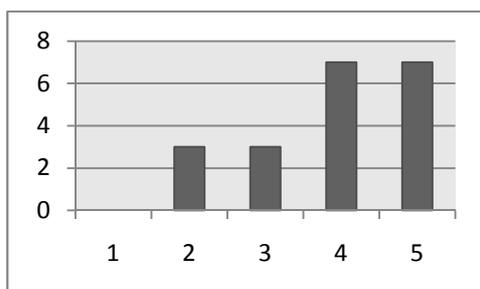


Table 3: Clarity of test utterance (4).

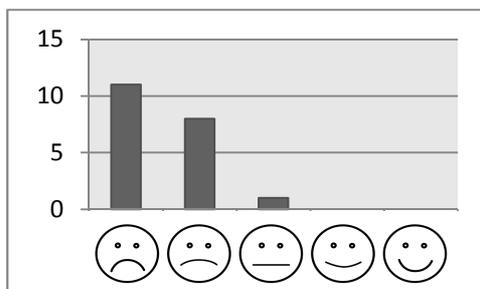


Table 4: Emotion of test utterance (4).

6. Conclusion and Future Work

The use of linguistic abstractions as a basis for animations has yielded promising results. The animations produced were well received by fluent signers and appear to communicate effectively. The data strongly suggest that the representation chosen for our corpus is flexible enough to display co-occurring facial nonmanual signals.

While this approach undoubtedly requires extension and revision, it is a step toward the automatic generation of American Sign Language. Moving forward, we plan to extend this representation to other parts of the body and test it with a wider range of utterances. We also plan to integrate the corpus structure into a more complete user interface that would facilitate the generation of ASL animations incorporating linguistic and extralinguistic features that interact on many levels including the facial nonmanual signals presented here.

7. Acknowledgements

We would like to acknowledge Nick Roessler and Brent Shiver for their help organizing and conducting user tests at DeafNation Expo, and Diana Gorman Jamrozik and Peter Cook of Columbia College Chicago for valuable discussions on nonmanual signals. We would also like to acknowledge DePaul University and The American Sign Language Project for funding.

8. References

- Brugman, H. & A. Russell (2004). Annotating Multi-media / Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation. IMDI Team (2003), IMDI Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen.
- Burtnyk, N. & Wein, M. (1976). Interactive Skeleton Techniques for Enhancing Motion Dynamics in Key Frame Animation. Communications of the Association for Computing Machinery, Vol 19. No. 10 October 1976, 564-569.
- DuCharme, B. (1999). XML: The Annotated Specification. Upper Saddle River, NJ: Prentice-Hall.
- Grossman, Ruth & Judy Kegl. 2006. To capture a face: a novel technique for the analysis and quantification of facial expressions in American Sign Language. Sign Language Studies 6(3) 273-305.
- Farhadi, A., Forsyth, D. & White, R. (2007). Transfer learning in sign language. In Computer Vision and Pattern Recognition, pages 1-8.
- Hanke, T. & Storz, J. (2008). iLEx – A database tool for integrating sign language corpus linguistics and sign language lexicography. In: Crasborn, Onno et al. (eds.): LREC 2008. 6th International Conference on Language Resources and Evaluation. Workshop Proceedings. W25. 3rd Workshop on the Representation and Processing of Sign Languages. Sunday 1st June 2008, Marrakech – Morocco. Paris: ELRA, 64-67.
- Reilly, J., McIntire, M. & Bellugi, U. (1990). Faces: The relationship between language and affect. In Virginia Volterra & Carol Erting (eds.), From Gesture to Language in Hearing and Deaf Children, New York, NY: Springer-Verlag. 128-141.
- Thomas, F., and Johnston, O. (1981). The Illusion of Life: Disney Animation. New York: Walt Disney Productions.
- Weast, T. (2008). Question in American Sign Language: A Quantitative analysis of raised and lowered eyebrows. PhD thesis, The University of Texas at Arlington.
- Wilbur R.B. (2000). Phonological and prosodic layering of nonmanuals in American Sign Language. In Lane, H. & K. Emmorey (eds.), The signs of language revisited: Festschrift for Ursula Bellugi and Edward Klima, (pp. 213-241) Hillsdale, NJ: Lawrence Erlbaum.
- Wolfe, R. (2006). An Improved Tool for Practicing Fingerspelling Recognition. Conference 2006 International Conference on Technology and Persons with Disabilities. Northridge, California, March 17-22.
- Wolfe, R. McDonald, J., Davidson, M., and Frank, C. (2007) Using an Animation-based Technology to Support Reading Curricula for Deaf Elementary Schoolchildren. The 22nd Annual International Technology & Persons with Disabilities Conference. Los Angeles, CA March 21.

Automatic sign language recognition: a social approach

Jesús Gumiel, Marina Serrano, José Miguel Moya

Telefónica I+D,

Granada, Spain

E-mail: jegumi@tid.es, marinas@tid.es, jmml@tid.es

Abstract

This paper reviews the social needs of the deaf community and describes the mechanisms and/or technologies which would improve the quality of life of this collective. The basis of this project is a teleinterpretation pilot, developed in Andalusia (Spain), and as a result of the interaction with the users, two investigation lines have been discovered, telephone communication, and e-learning. These activities have a clearly defined technology need by the hearing impaired, and existing solutions do not completely solve the problem, therefore they are a good scenario to implement an automatic sign language recognition system. The aim of the paper is to demonstrate how to thanks to this technology, social barriers can be torn down, allowing equal access to those services that today are restrictive for deaf people.

1. Introduction

In a multicultural world where people are in constant movement and distance is no longer a problem, the definition of the word ubiquitous has increased its meaning. In this context, it is of vital importance to ensure that the groups at risk of exclusion, including deaf people, have access to new communication technologies.

By developing the appropriate tools, such as automatic sign language recognition systems, these technologies can promote the elimination of existing communication barriers, providing mechanisms for social integration.

With this target in mind, the platform of communication for deaf people was born; not only as a tool to resolve the problems of distance, but also to improve the quality of life of the deaf, adding new and promising possibilities to their standard methods of learning, signing, and interaction with hearing people.

The platform has two main focuses: on one hand it would constitute help to the traditional system of sign language interpretation, in cases where an interpreter is unable to reach the place where the deaf person is, or even when the demand for interpreters supersedes the number on hand.

On the other hand we have the learning of sign language: reducing the difficulties to link a SL signal with a written word and promoting the early communication in an effective way between hearing parents and deaf children, allowing the linguistic, psychological and social development of these children to be easier.

The awareness of the importance of sign language and the methods used today to teach it is an important research issue in Europe and this is where the platform must demonstrate its advantages and provide its services, not only to a individual but also for groups of e-learners.

This system relies on new communications research, such as New Generation Intelligent Networks (NGIN), Voice over IP (VoIP), advanced videoconferencing techniques and network integration (3G, analogical telephony, IP telephony, internet applications).

The addition of an automatic sign language recognition system will allow the deployment of a variety of unattended SL services, easily accessed. This tool will play an important role in the current technological environment, disseminating the use and knowledge of sign language and encouraging people participation.

1.1 Spanish Sign Language

Spanish Sign Language, being the languages of the deaf and deaf-blind who have opted for this modality linguistics, have not had the recognition, nor proper development, and despite the fact that numerous investigations carried out nationally and internationally have shown that sign languages meet all the requirements of natural language and possess a grammatical syntaxis and lexical features of its own. Recently this situation has been corrected and proof is the adoption of many standards, including most notably several Statutes of Autonomy, who recognize the importance of sign languages.

One example of the importance of sign language is the number of users of sign language in Spain, approximately 400,000 of which 100,000 are deaf. (CNSE, 2010).

Sign language is the main pillar on which underpinned services, created for the ease of social integration of deaf people, as it will be the medium used for convergence communication between users of services, both hearing users and hearing impaired.

1.2 Interpretation

In Spain there are a total of 2,781 sign language interpreters, accredited professional training among non-formal and formal training graduates, of which about 25.17 percent are active.

According to this data, in Spain, the ratio of sign language interpreters there is a professional for every 143 people who are deaf or hearing impaired (SID, 2010).

Recently, the Federations of the National Confederation of Spanish Deaf (CNSE) voiced the need to incorporate sign language interpreters into public life, the creation of the Center for Linguistic Normalization of Spanish Sign Language and to promote learning. These claims are part of the manifesto drawn up for the celebration, last September, the International Day of Deaf Persons.

The confederation requires full accessibility to the audiovisual contents through subtitling and to incorporate content broadcast in sign language, so therefore "are guaranteed the rights of all deaf people to receive information, for the enjoyment of leisure and culture".

1.3 Deaf Community in Spain

The Deaf Community is endowed with an associative structure with dense networks of relationships, organized around institutions and distinctive culture. Culture in the double sense of belief systems, values, shared practices and cultural productions such as narrative, storytelling, humor, puns, sign language poetry, drama and mime, sculpture, painting, photography and films sensitive to the experiences of deaf people.

It is a living community, varied and open to all sorts of people whose central element is sign language. In Spain, the law 27/2007 of 23 October, mentions the "linguistic community of the people who use sign language Spanish (BOE, 2007).

	+6 ages	6-64 ages	64+80 ages
Disabilities	Total	Total	Total
Listen	961.489	295868	665621
Disability to receive any sound	102.394	46952	54442
Disability for hearing loud sounds	230.736	64906	164830
Disability listening to the speech	815.639	234164	581745
Communicate	504.813	244546	260267
Communicating through speech	173.449	71141	102308
Communicating through alternatives language	88.642	50813	37829
Communicating through unsigned gestures	69.765	33739	36026
Communicating through writing / reading conventional	414.981	191886	223095

Figure 1: Table numbers in communicating disabilities (INE, 2010)

2. Pilot of TeleInterpretation Center

In January 2009 in Granada (Andalusia, Spain) an innovating project created by Telefónica in collaboration with the FAAS (Andalusian Federation of Associations of the Deaf) and promoted by the Andalusia local government. The target of this project was the development of a centre of teleinterpretation where the users can access telephone services through interpreters.

Deaf people find communication barriers in the access to some services of Public Administration that can be done by phone, such as:

- Request information.
- Emergency calls.
- Request an appointment.

With the target of making easy the communication between deaf people and the public entities in Andalusia, a collaboration agreement was created between different entities. As a result of this one started the LSE (Spanish Sign Language) teleinterpretation service.

2.1 The way the service works

Deaf people have a videotelephone call to the center and request the interpreter to make a call to a public entity.

By this way the contact between deaf people and hearing person is established.

This flow of communication is shown in the next figure.



Figure 2: Schema Teleinterpretation Centre

2.2 Information and statistics

The pilot has been working for 1 year, with 2 interpreters and around 20 deaf users making an average of 80 calls per month.

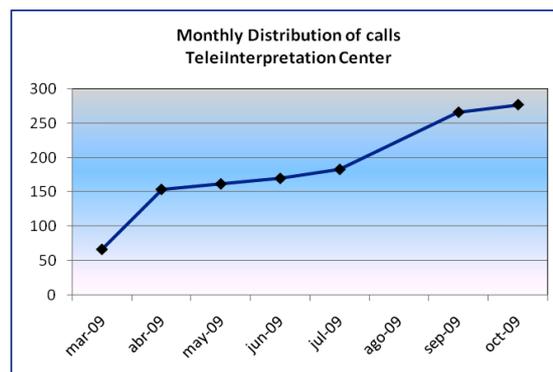


Figure 3: Statics of Teleinterpretation Centre

At first the timetable of the center was only in the morning, but because of the petitions of the users this had to be increased to morning and afternoon.

Another important feature of the pilot is that the users can make call to other deaf persons directly, without the intervention of an interpreter. So the center not only communicates deaf people with administration and other services, but it also allow users to make personals video calls between their, improving highly their communication.

One interesting fact is that the services most in demand are; asking for gas cylinders, communicating with their lawyer, the plumber..., something usual for everybody, but a complete novelty for hearing impaired users.

2.3 Results and improvements

Thanks to the experience obtained in the pilot, both, social and technological, some basic points have been defined with regards to the new development in the area of heading impaired. Here some of the next ones:

- There are not enough interpreters to attend the requirements of deaf users. Even using the platform sometimes there are users waiting.

- It is necessary develop an automatic system that can be used when the use of an interpreter is not possible.
- The system of subtitles is not a global solution. There are an important number of illiterate deaf people because of the problems with learning, and babies that are learning Sign Languages do not know how to read yet.

3. The SignSpeak Project

The SignSpeak project is one of the first EU funded projects that tackles the problem of automatic recognition and translation of continuous sign language.

The overall goal of the SignSpeak project is to develop a new vision-based technology for recognizing and translating continuous sign language (i.e. provide Video-to-Text technologies), in order to provide new e-Services to the deaf community and to improve their communication with the hearing people (Dreuw & Ney & Martinez & Crasborn & Piater & Moya & Wheatley, 2010).

The current rapid development of sign language research is partly due to advances in technology, including of course the spread of Internet, but especially the advance of computer technology enabling the use of digital video (Crasborn et al., 2007). The main research goals are related to a better scientific understanding and vision-based technological development for continuous sign language recognition and translation:

- Understanding sign language requires better linguistic knowledge
- Recognition of large vocabularies requires a more robust feature extraction methods and a modeling of the signs at a sub-word unit level
- Statistical machine translation requires large bilingual annotated corpora and a better linguistic knowledge for phrase-based modeling and alignment.

Therefore, the SignSpeak project combines innovative scientific theory and vision-based technology development by gathering novel linguistic research and the most advanced techniques in image analysis, automatic speech recognition (ASR) and statistical machine translation (SMT) within a common framework.

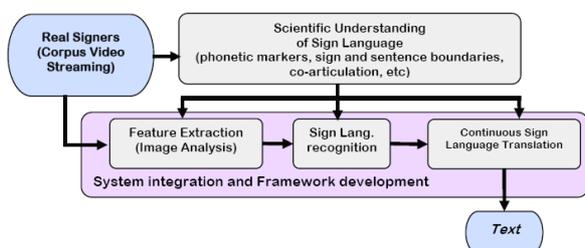


Figure 4: Conceptual scheme of SignSpeak

4. Automatic Sign Language Recognition

Nowadays one of the main problems for deaf people is the lack of interpreters. Even with system such as the platform of teleinterpretation commented on in point 2 of this paper, the number of interpreters is still insufficient for the needs of the deaf community.

One of the technologies which can propose a solution to

this problem is the automatic sign language recognition (ASR), what is the conversion of a signal into a sequence of written words.

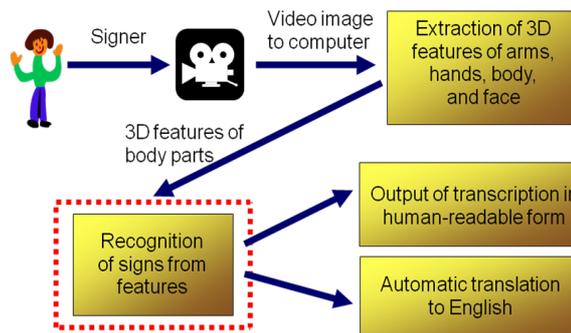


Figure 5: Schema of ASL Recognition

The ASLR would be the entrance to the system, and would allow the user communicate in real-time with a hearing person without the need for an interpreter. Another advantage to this type of application is the Internationalization, it means, the user could use the Spanish Sign Language for example, and the result of the translation could be an English text. It would be question of choose the correct parameters in the platform.

ASLR will be present in the future, and some previous problems with this technology are disappearing thanks to the advance of other researches related to this, such as gesture recognition, recognition of human actions, etc. Another reason that assures the accuracy of ASLR is the appearance of linguists in this field of investigation, and the interest of this in helping with the data translation.

In spite of the advance of the research in ASLR, this still remains a complex technique with various problems and limitations, but the target of this paper it is not to focus in the way of a system that can recognize sign language, but also to propose uses of this hypothetical futures system that improving quality of life of deaf people.

5. Future Applications based in ASLR

Once we have explained the two projects that have been served as the basis of the research about deaf community requirements, the next step is to develop the future applications that help to resolve the communication problems of hearing impaired.

These new applications will be focusing in two areas very representative of the collective, tele-interpretation and e-learning.

5.1 Teleinterpretation

Thanks to experience obtained with the teleinterpretation center pilot in Andalusia, Spain, a great number of features has been listed, that according to the users will improve the quality of this kind of service.

It is important to highlight the benefits of the teleinterpretation, which allow the user carry out tasks from home using the phone. This is a benefit for the all deaf community, because it maximizes the services that one interpreter can offer, and avoids travel time and expenses.

Even with teleinterpretation the number of

interpreters is not enough for the quantity of users, and in emergency situations is not possible to wait in a queue for your turn, so this service is one perfect candidate to make use of the ASLR.



Figure 5: Too much users per interpreter

With the ASLR the users call to the center, and they can choose if they want to talk to an interpreter or talk directly with the ASLR system, in this case the application would translate, real-time the signs of the user to a text, which could be converted to an audio message by a speech tool or sent by mail or sms to the receiver.

The waiting would disappear, and the communication would be fluent and without latency. The wish of the majority of users would be fulfilled.

5.2 E-learning

The main request of the deaf associations is a system to improve the method of learning sign language.

Regarding the results and researches carried out both in teleinterpretation Center and in the SignSpeakers project; the deaf community has some great problems when teaching:

- Students have difficulties to relate concepts with signs.
- It is very uncomfortable for students, above all, for those than live in a rural area, travel to the teaching center.
- It is difficult to create a group of younger students with the same level.

The solution to all these problems is the creation of an e-learning system that give the user the possibility, of connecting from home to a learning room no matter where the lessons is being provided. The teacher explains the concepts in an e-learning center, and his signs will be translated to text and also there will be images and videos added to the explanation to make easy the understanding of the class. This process is developing in real-time, with the interaction of all the participants.

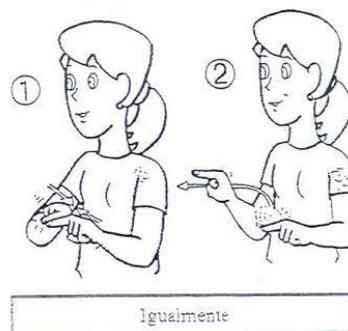


Figure 6: Learning with videos and images

6. Conclusions

Technology can improve the life of deaf people, but it is necessary to research the real need of the collective and the best way to get close the new tools to the users. Without the collaboration of the deaf groups probably this improvements would not be taken into account by the community.

In future when these technologies mature, they may help improve the independence of deaf people because in certain situations, they will not need an interpreter to accompany them.

7. Acknowledgements

8. This work received funding from the European Community's Seventh Framework Programme under grant agreement number 231424 (FP7-ICT-2007-3).

9. References

- BOE, Gazette, Spain, 2007, <http://www.boe.es/boe/dias/2007/10/24/pdfs/A43251-43259.pdf>
- CNSE, Confederation of deaf people, 2010 Spain, <http://www.cnse.es/>
- Christian Vogler, Automated Sign Language Recognition, Gallaudet Research Institute, Gallaudet University. Washington, DC 20002-3695, USA <http://gri.gallaudet.edu/~cvogler/>
- INE, National Statistical Institute, Spain, 2010, http://www.ine.es/prodyser/pubweb/disc_inf05/disc_a_ig_cap2.pdf
- Mark Wheatley and Annika Pabsch. 2010. Sign language in Europe, this volume.
- Philippe Dreuw, Hermann Ney, Gregorio Martinez, Onno Crasborn, Justus Piater, Jose Miguel Moya, and Mark Wheatley. 2010, The SignSpeak Project - Bridging the Gap Between Signers and Speakers, this volume.
- SID, Disability Information Service, 2010, Spain, <http://sid.usal.es/noticias/discapacidad/35003/1-1/es-pana-tiene-2781-interpretes-de-lengua-de-signos-un-o-por-cada-143-personas-sordas.aspx>

RWTH-Phoenix: Analysis of the German Sign Language Corpus

Daniel Stein, Jens Forster, Uwe Zelle, Philippe Dreuw, and Hermann Ney

Human Language Technology and Pattern Recognition
RWTH Aachen University, Germany
surname@cs.rwth-aachen.de

Abstract

In this work, the recent additions to the RWTH-Phoenix corpus, a data collection of interpreted news announcement, are analysed. The corpus features videos, gloss annotation of German Sign Language and transcriptions of spoken German. The annotation procedure is reported, and the corpus statistics are discussed. We present automatic machine translation results for both directions, and discuss syntactically motivated enhancements.

1. Introduction

For data-driven automatic sign language processing, finding a suitable corpus is still one of the main obstacles. Most available data collections focus on linguistic issues and have a domain that is too broad to be suitable for these approaches. In (Bungeroth et al., 2006), the RWTH-Phoenix corpus was described, a collection of richly annotated video data from the domain of German weather forecasting. It includes a bilingual text-based sentence corpus and a collection of monolingual data of the German sentences. This domain was chosen since it is easily extendable, has a limited vocabulary and features real-life data rather than material made under lab conditions.

In this work, we are going to analyse the recent additions made to the existing corpus and its impact on the automatic machine translation. We are also applying some recent advancements in the field of statistical machine translations and analyse if they work on tiny data collections.

1.1. Related Work

Recently, a couple of other sign language data collections have been created. Based on their initial purpose, some of them have only limited usability to data-driven natural language processing techniques. Listed below are some of the larger efforts for European sign languages.

ECHO The European Cultural Heritage Online organization (ECHO)¹ published data collections for Swedish Sign Language, British Sign Language and Sign Language of the Netherlands. Their broad domain of children's fairy tales as well as poetry make them rather unsuitable for statistical methods. Another obstacle is the intensive usage of signed classifiers because of the rather visual topics.

Corpus NGT (Crasborn and Zwitserlood, 2008) present a data collection in the Sign Language of the Netherlands. It consists of recordings in the domain of fables, cartoon paraphrases, discussions on sign language and discussions on Deaf² issues. In the european funded

Signspeak project³, sentence-aligned translations into Spoken Dutch are currently ongoing.

ATIS In (Bungeroth et al., 2008), a corpus for English, German, Irish Sign Language, German Sign Language and South African Sign Language in the domain of the Air Travel Information System (ATIS) is given. With roughly 600 parallel sentences in total, it is small in size. However, being a multilingual data selection, it enables direct translation between sign languages.

Czech-Signed Speech In (Kanis and Müller, 2009), a data collection for Czech and Signed Czech is presented. Its domain is taken from transcribed train timetable dialogues and then translated by human experts. However, the actual translations are not in the Czech Sign Language spoken by the Deaf, but in an artificial language system strongly derived from spoken Czech. Explicit word alignments are made by human experts. Due to its nature, the authors are able to achieve with very high performance scores.

1.2. Paper Structure

This paper is organised as follows. We analyse the current status of our data collection in Section 2., with special attention to the transcription process and the corpus statistics. In Section 3., the translation methods and results are presented, including syntactically motivated enhancements to the translation system. In Section 4., a summary and an outlook are given.

2. Corpus Analysis

The public broadcast channel "Phoenix" offers live interpretation into German Sign Language (DGS)⁴ for the main evening broadcast news. Its videos are recorded automatically by our servers.

Since the last batch of recordings in 2005 (Bungeroth et al., 2006), the television program has changed in two important aspects. First, the format of the video is different: before, the news announcer was slightly distorted in perspective, and the signing interpreter was shown without a background of its own. Now, the broadcast channel shows

¹<http://www.let.kun.nl/sign-lang/echo>

²Following common conventions, we denote the cultural group of deaf people with a capital "D"

³<http://www.signspeak.eu/>

⁴Deutsche Gebärdensprache

the original video in a smaller frame and places the signing interpreter in front of a grey background on the far right (cf. Figure 1). For machine translation it does not pose a problem since the algorithms only work on the transcriptions and not on the video signal.

As for the second major change in the data, the transcription of the audio material is no longer provided by the broadcast station. We therefore employ an automatic speech recognition system for the German audio data which transcribes the spoken words, and manually align the words to the annotated gloss sentences. For the weather forecast, the audio recognition word error rate is well below 5%, making the transcription quite convenient.

2.1. Quality and Usability

Although interpreted by bilingual experts, the translation into German Sign Language quality suffers from the recording situation: the interpreters have to listen to the text under real-time conditions and thus have to sign simultaneously without preparations. Due to the complex nature of official news announcements and the relative speed of the announcer, the signed sentences are still in German Sign Language but tend to have a slight bias towards the grammar structure of spoken German. Also, details are omitted in the signed sentences. For example, if the temperature for the region of Bavaria, the adjacent Austrian Alps and the river Donau is described in the weather forecast, the interpreter might refer more generally to the south of Germany without specifically naming the exact locations. Another typical omission occurs when the announcement refers to specific wind velocities such as “schwach”, “mäßig”, and “frisch” (being a 3, 4 and 5 on the Beaufort scale, respectively), the interpreters typically only differentiate between a low and a high velocity.

The notion of a signed sentence is an active research topic in the linguistic community. Here, we take a rather pragmatic (and probably erroneous) approach and match the gloss output to the spoken German sentences, i.e. we split gloss sentences transcribed by our deaf colleague if their topic stretches over more than one German sentence. In a second-pass, we also omit all information in the spoken German sentences that are clearly not signed by the interpreter, but try to stay as close to the previous grammar structure as possible.

2.2. Notation

According to common conventions, glosses are generally written in upper case. Incorporations are treated as a single word, finger-spelled words and compound words are joined by a +. Dialectal forms are stored in a simple database so that they are mapped to the same word for translation but appear differently for the recognition (e.g. “WOMAN1” for the Bavarian sign for woman and “WOMAN2” for the dialectal form used in the northern part of Germany). If a sign is repeated fast and without a specific number, a double + is written at the end of the sign (e.g. “ASK++”, which translates to *enquire* rather than *asking*). If a sign is repeated a specific number of times to mark multiple occurrences, they are denoted separately (e.g. two groups of clouds are denoted as “CLOUD CLOUD”). Additional information that

carries crucial semantic information is denoted as:

- loc:** for a specific location with a spatial reference (e.g. “loc:coast” for the coast in the northern part of Germany, but also “loc:from_north_to_south” for a southward movement)
- mb:** mouthing that is important to discriminate the word meaning, (e.g. “RIVER-(mb:rhein)” and “RIVER-(mb:donau)” for the different rivers which have the same manual movement)

Apart from this, we annotate hand movement not related to a signed word. $\langle \text{ON} \rangle$, $\langle \text{OFF} \rangle$ is used for signing onset and offset, $\langle \text{PU} \rangle$ is a palm-up gesture, and $\langle \text{EMP} \rangle$ marks emphatic movement that is not a sign (e.g. when the interpreter is shrugging the shoulder). For the translation experiments below, we treated the mouthing and location information as normal words.

2.3. Annotation

For the annotation, we made use of the free ELAN tool developed at the Max Planck Institute for Psycholinguistics in Nijmegen⁵. Start and end times are marked on a sentence level rather than on the word level. Both left-hand and right-hand movements are kept track of independently.

Our annotator is congenitally deaf and has worked in research fields regarding sign language for over a decade, but had no previous annotation experiences. According to his feedback, it took him about two weeks to get accustomed to the annotation tool. For the first two month working on the recordings, there were various questions coming up about the annotation procedure, namely for such effects as dialects, synonyms, classifiers, left-hand/right-hand issues which were discussed in his mother tongue with interpreters. At first, it took him 4 hours for one weather forecast of roughly one minute. After two months, he was able to finish three videos in the same time amount. For the whole news announcement, which has a basically unlimited domain and runs for 15 minutes, it takes him about 24 working hours to transcribe it.

2.4. Corpus Progression

In an ongoing process, the corpus has recently been extended with additional material. For the transcription of the glosses and their translation into spoken German, they blend in with the old annotations and can be used together for statistical machine translation. So far, 43 new videos were added to the existing 78 videos.

Comparing the corpus statistics with other small-sized data selections, the domain seems to be suitable. For example, the Chinese-English task of the International Workshop on Spoken Language Technology (IWSLT)⁶, is a selection of parallel sentences in the domain of travel and booking information, has 22 K training sentences, with a token-type ratio of 18.8 for Chinese and 27.5 for English. Compared to our corpus, we currently have a total of 2.7 K training sentences and already approach a type-token ratio of around

⁵<http://www.lat-mpi.eu/tools/elan/>

⁶<http://mastarj.nict.go.jp/IWSLT2009/>



Figure 1: Old and new television format used in the Phoenix television channel

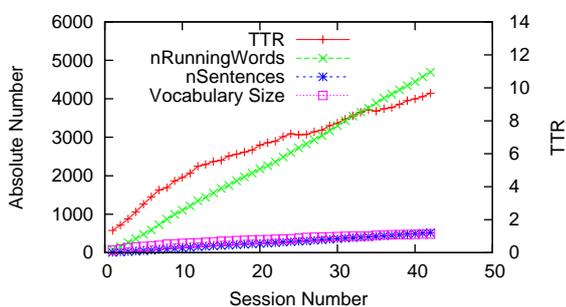


Figure 2: Number of sentences, vocabulary size, type-token ratio, for the newly annotated data

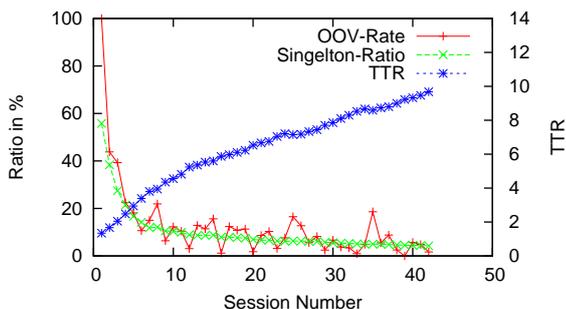


Figure 3: Out of vocabulary and singletons for the newly annotated data

10 (Figure 2 and 3) after 40 sessions. The singleton ratio is about 40% for both languages in IWSLT, while ours goes quickly below 20% and stays there. The peaks in singletons and out-of-vocabulary ratios can mostly be attributed to time-specific terms like the easter season or certain places where weather phenomena occur in a certain week. Since these words tend to occur often in consecutive sessions, the singleton ratio typically drops fast. For a complete corpus overview, see Table 1.

3. Translation

We use an in-house statistical translation system similar to (Chiang, 2005). It is able to process hierarchical phrases in a context-free grammar with a variation of the CYK algorithm. For a given sentence f , the best translation \hat{e} is chosen as the target sentence e that maximizes the sum over m different models h_m , scaled by the factors λ_m :

$$\hat{e} := \operatorname{argmax}_e \left(\sum_m \lambda_m h_m(e, f) \right). \quad (1)$$

The alignment is created for both translation directions with GIZA++⁷ and merged with a variation of the growdiag-final algorithm. We employ a trigram language model using modified Kneser-Ney discounting which is trained with the SRI toolkit⁸. The scaling factors of the log-linear model are optimized on the development set with Och's Minimum Error Rate Training (Och, 2003), which is a variation of Powell's method working on n -best translations. The resulting factors are then used to translate the test set. For automatic error measures, we use the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2001), which is based on n -gram precision and has a brevity penalty for sentences that are too short. Further, we use the Translation Error Rate (TER) (Snover et al., 2006), which is similar to the Levenshtein distance but allows for shifts of word blocks. Note that BLEU is better if higher and TER is better if lower.

In order to enhance the statistic reliability of the results, we opted to increase the number of sentences withheld from the training material for development and test set to 20% rather than 10% in our previous publications. Further, cross-validation has been carried out, taking three different splits of the data into the training, development and testing set, with completely independent alignment creation, language model and optimization. The results between the splits are not comparable in this way, but a consistent improvement in all splits backs up the usefulness of the applied method.

⁷<http://www.htlpr.rwth-aachen.de/~och/software/GIZA++.html>

⁸<http://www-speech.sri.com/projects/srilm/>

		DGS	German	preprocessed German
train	#sentences	2711		
	#running words	15499	21679	22891
	vocabulary size	916	1476	1180
	#singletons	337	633	434
dev	#sentences	338		
	#running words	1924	2689	2832
	#OOVs	33	65	50
	#sentences	338		
train	#sentences	338		
	#running words	1750	2629	2773
	#OOVs	48	49	32

Table 1: Corpus overview of the RWTH-Phoenix corpus for one specific split of the data. The numbers are similar for the other two splits. The preprocessing of German is explained in Section 3.1.

3.1. German to Glosses

Sign languages lack a formal written form universally accepted by the Deaf. Thus, gloss annotations are typically only employed by linguistic experts, but they can be used to feed avatars with signing input. Being single-reference experiments, the quality of the output is reasonable but not without flaws. Looking at the examples in Table 3, we can see that the translation system was able to come up with some of the typical reorderings taking place in the grammar of the two languages, but failing to translate words that are highly flexed in German and thus lead to data sparseness problems.

We therefore reduced the morphologic complexity of the German source language by automatic means. To achieve this, we parsed the data with a morpho-syntactic analysis tool before the actual translation. The freely available tool Morphisto⁹ is a finite-state transducer with a large database of German, accurately reporting part-of-speech tags, gender, casus and possible split points for large compound words. However, if ambiguous it does not provide probability scores for the various possible parsings. We therefore opted to always take the entry consisting of the fewest split points possible (cf. Table 2). By doing so, we reduce all words to their stem form and split large words automatically. In (Stein et al., 2006), it was already shown that these methods help enhance the translation quality.

In Figure 4, an example for an improvement in alignment quality is given. In Table 5, the results for this task are presented.

3.2. Glosses to German

This translation direction is more challenging since the German announcements often appear to be more varied and even lyrical in nature. Even though the interpreter always speaks of a clear sky during the night (“HEUTE NACHT KLAR”), the announcer will sometimes refer to the dissolving of the clouds, a clear sky or the sparkling of the stars. We are not able to preprocess the input automatically since no morpho-syntactic parser for the glosses exist, and a reduction of the target language complexity dur-

input	Wettervorhersage
	wetten-er-Vorhersage
	wettern-Vorhersage
	Wetter-Vorhersage
	...
output	wettern Vorhersage

Table 2: Different breaking points proposed in Morphisto for the German word “Wettervorhersage” (english: weather forecasting). The last one is correct, but the second is taken in our heuristic.

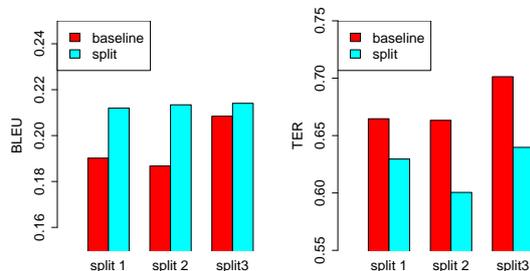


Figure 5: BLEU and TER results for German to German Sign Language Translation on three different test sets

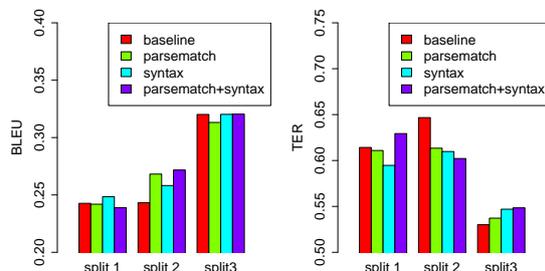


Figure 6: BLEU and TER results for German Sign Language to German Translation on three different test sets

ing translation would require a rather sophisticated post-processing that possibly introduces further errors.

However, we can make use of some syntactic analysis of the target language and enforce the structure of the German grammar onto our decoder. In this work, we opted for two methods. The first measures the compatibility of the phrases with a node in a deep syntactic tree, preferring complete sub-sentence structures such as noun phrases or verb phrases. If the target phrase does not match a node, we take the minimal amount of words needed to reach a fitting node as penalty, similar to (Vilar et al., 2008). We denote these experiments as *parsematch*.

Also, we employ soft syntactic features as in (Venugopal et al., 2009). With this, we replace the generic non-terminal label used in common hierarchical decoding and replace it

⁹<http://code.google.com/p/morphisto/>

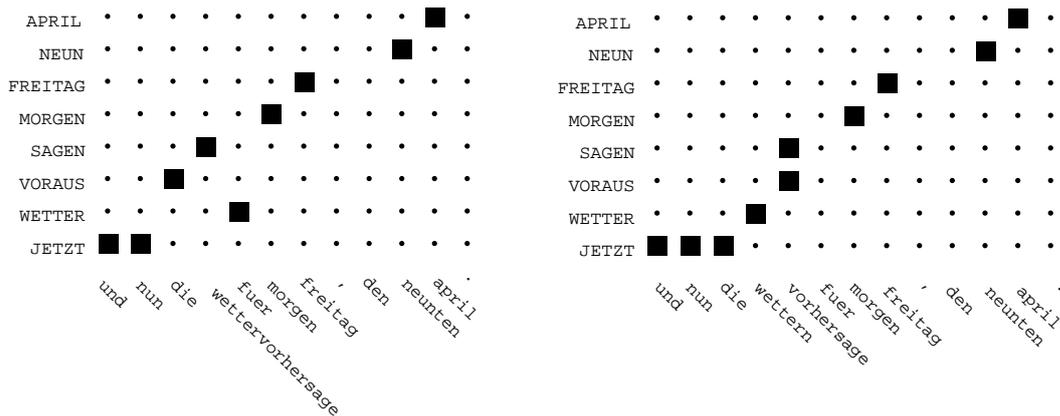


Figure 4: Alignment before and after splitting. The left one is more accurate.

source	Im Norden fällt etwas Regen bei stürmischem Wind .
baseline translation	NORDEN BISSCHEN REGEN HIER unknown_stürmischem
split translation	NORDEN BISSCHEN REGEN STURM-emp
reference	NORDEN WENIG REGEN STURM-emp
source	Diese Regenwolken ziehen heute Nacht aus Frankreich heran .
baseline translation	IX REGEN ZIEHEN HEUTE NACHT FRANKREICH
split translation	AUCH REGEN WOLKE ZIEHEN HEUTE NACHT FRANKREICH ZIEHEN-(loc:nach_mitte)
reference	REGEN WOLKE ZIEHEN++ FRANKREICH IX WOLKEN ZIEHEN-(loc:nach_mitte)

Table 3: Translation examples for German to German Sign Language

with phrase tags from the syntactic parser. Thus, we now have a variety of 65 non-terminals and see if a new translation matches the syntactic label that it tries to replace. This is denoted as *syntax*.

Note that we do not restrict the regular translation by doing so but merely offer another translation model to the log-linear model, thus theoretically allowing the decoding process to ignore it by setting the according scaling factor to 0. The parsing was done using the freely available Stanford parser¹⁰.

In Figure 6, the results for this task are presented, with some examples in Table 4. While in general the BLEU score improved in all development optimizations, the results on our test test were not consistent. Possible reasons for this were the large number of labels that the Stanford parser produces, compared to the small data set. In a next step, we plan to reduce their number by means of automatic clustering. We also noted an increase in the TER score on some tasks, possibly by enforcing larger phrases with the syntactic models.

4. Conclusion

We presented and analysed the recent extensions to the signed weather forecasting corpus RWTH-Phoenix and tested various syntactically motivated methods to enhance the statistical machine translation on this task. It is currently one of the largest data collections for a natural sign language and designed for the needs of statistical translation and recognition. Great care has been taken to ensure

that all the above mentioned methods and tools are freely available to the scientific community. Also, our complete hierarchical translation system will be released as open source in the near future.

The data collection is available upon request. We hope that the performance on this task can be taken for comparison and serve as a benchmark for other groups working in this field. As an outlook, we look forward to combine the recognition results made on the videos with our translation system. We also hope to further incorporate additional recognition input such as tracking of the hand position into our translation models.

5. Acknowledgments

This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics - STREP) under grant agreement n° 231424.

6. References

- Jan Bungeroth, Daniel Stein, Philippe Dreuw, Morteza Zahedi, and Hermann Ney. 2006. A german sign language corpus of the domain weather report. In *International Conference on Language Resources and Evaluation*, pages 2000–2003, Genoa, Italy, May.
- Jan Bungeroth, Daniel Stein, Philippe Dreuw, Hermann Ney, Sara Morrissey, Andy Way, and Lynette van Zijl. 2008. The atis sign language corpus. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

¹⁰<http://nlp.stanford.edu/software/lex-parser.shtml>

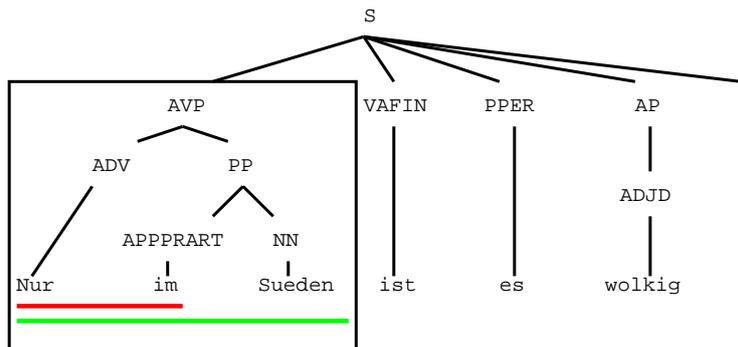


Figure 7: Stanford parsing of a German sentence. A phrase consisting of “Nur im” would be considered bad since it does not match a parse tree node. A good phrase would be “Nur im Sueden”. For syntactic labeling, both phrases would be mapped to the same node, i.e. “AVP”.

source	KALT LUFT MITTELMEER ZIEHEN-(loc:nach_mitte)
baseline	Kalte Mittelmeer Luft zu uns .
phrasematch	Kalte Luft Mittelmeer zieht zu uns .
reference	Kalte Mittelmeerluft zieht nach Deutschland .
source	DA TIEF DRUCK LOC REGION WESTEN DRUCK
baseline	Dort ein Tiefdruckgebiet liegt im Westen Hochdruckzone .
syntax	Weitgehend ein Tiefdruckgebiet im Nordwesten Tiefdruckgebiet bestimmt .
reference	Im Westen liegt ein Tiefdruckgebiet .

Table 4: Translation examples for German Sign Language to German

- David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, Michigan, June.
- Onno Crasborn and Inge Zwisserlood. 2008. The Corpus NGT: An Online Corpus for Professionals and Laymen. In Crasborn, Hanke, Efthimiou, Zwisserlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages at LREC 2008*, pages 44–49, Paris. ELDA.
- Jakub Kanis and Luděk Müller. 2009. Advances in czech signed speech translation. In *Lecture Notes in Computer Science*, volume 5729, pages 48–55. Springer.
- Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.
- Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. pages 223–231, Cambridge, MA, August.
- Daniel Stein, Jan Bungeroth, and Hermann Ney. 2006. Morpho-syntax based statistical methods for sign language translation. In *Conference of the European Association for Machine Translation*, pages 169–177, Oslo, Norway, June.
- Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference Grammars: Softening Syntactic Constraints to Improve Statistical Machine Translation. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–244, Boulder, Colorado, June.
- David Vilar, Daniel Stein, and Hermann Ney. 2008. Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation. In *International Workshop on Spoken Language Translation*, pages 190–197, Waikiki, Hawaii, October.

Development of E-Learning Service of Computer Assisted Sign Language Learning: Online Version of CASLL

Saori Tanaka[†], Yosuke Matsusaka[‡], Kaoru Nakazono^{††}

[†]MID
Nonoshita, Nagareyama, Chiba,
270-0135, Japan
tanaka@mid-japan.com

[‡]National Institute of Advanced Industrial
Science and Technology
Umezono, Tsukuba, Ibaraki 305-8568, Japan
yosuke.matsusaka@aist.go.jp

^{††}NTT Network Innovation Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo,
180-0012, Japan
nakazono@core.ecl.net

Abstract

In this study, we introduce the problems for realizing an e-learning system available online and outline some ethical issues behind these problems. The difficulties faced to us, when we were going to open Computer Assisted Sign language Learning (CASLL) system online, were one to expose the sign language movies to public with downloadable way, one for increasing the course materials, and one to enhance the collaboration between learners. The ethical discussions revealed that the reliability for the system and the collaborative work for expand the number of course materials were necessary for overcoming the difficulties. In order to realize the reliable and the collaborative e-learning system, we implemented CASLL within Moodle, an open-source Course Management System. For re-designing the system to actual use for sign language learners and teachers, we added new functions to Moodle; the protection function for the right of publicity, the wiki function to enable collaborative course editing and finally the Link function to enhance public relations. We are going to evaluate the system design from the view point of the usability for teaching, the effectivity for learning, and the utility for collaboration.

1. Introduction

In recent years there are strong demands for sign language teaching and learning. Computer Aided Education (CAE) of sign language is considered as one of the most effective way to assist teaching and learning activity of sign languages.

In the previous researches, technology that has used sign language can be classified into two groups of different aims. One group is the research aimed to support the deaf people themselves, in order to fill the social gaps between the deaf people and the hearing people. The other the research aimed to disseminate the knowledge of sign language in a society, so that the circumstances around the deaf people are improved.

In the first group for supporting deaf people themselves, the research for building animation generation system is in progress for Greek sign language (Efthimiou et al., 2004). This system is aimed to assist deaf people by converting the spoken language to the sign language. In Japan, based on the knowledge that the speed for playing JSL movies depends on the level of proficiency for JSL, a system for playing JSL movies in five speed level has been developed (Isono et al., 2006) to support the deaf people who has difficulty to read the sign language in fast speed.

In the second group for supporting disseminate the sign language knowledge, a remote communications system to connect a class of American sign language has been developed (Lehman and Conceicao, 2001) to assist the learning activity of the student who learn in different sites. In Japan, a learning system of finger spelling with feedback function (Tabata et al., 2001) has been developed for self learning activity. There is also the JSL database with search function based on the linguistic knowledge of native Japanese signers (Fukuda, 2005), and the video teaching material of JSL to assist teaching activity of the sign language (for JSL Learning,). The current situation for the

second group of research is that there is no e-learning system where the learners can learn JSL by themselves and collaborate with other JSL learners in remote areas at the same time.

Our study can be categorized in the second group of research for supporting disseminate the sign language knowledge. In the previous research, we have proposed a new learning program CASLL (Computer Assisted Sign Language Learning system) (Tanaka et al., 2007a) and compared to existing learning systems, as one of our series of studies for developing human interface by using JSL contents (Tanaka et al., 2007b; Nakazono and Tanaka, 2008; Tanaka et al., 2008). The system show some effectiveness, but we also have found new problems we did not expected (explained in Section 2.1.). The aim of this research is to improve the CASLL system by introducing the new design.

2. New System Design of CASLL

2.1. Problems in the Previous Version of CASLL

After we have started the development for online version of CASLL system in 2008, we encountered following difficulties in actual operation:

Difficulty to expose the movies to public In the previous version of CASLL, there are no function protect downloading the movie materials. Many signers resisted to expose the movies which captured their faces to online in downloadable way.

Difficulty for increasing the course materials In the previous version of CASLL, teacher had to edit the course in script form using a text editor. There was no easy interface that the JSL teachers who does not have knowledge on the script could edit the course.

Difficulty to enhance collaboration between the learners

In the previous version of CASLL, there was a function to give a feedback about the answers, but there

was no way to give feedback from the teacher or collaborative between the learners. Such collaboration is essential to make learning activity effective.

2.2. Ethical Background of the Problems

To understand the cause of above problems, we believe, not only the consideration for the technical issues of the system, but also the one for some ethical backgrounds of the training target is crucial.

Since 2004, the academic societies in Japan, especially for the engineering societies, researchers have been more strictly required to discuss about the research ethics and inform the purpose and risk of their research in easy-to-understand language to ordinary people (S. Tanaka, 2009b). As part of this movement, the Japanese Association of Sign Language Linguistics (JASL) also held a symposium which theme is “Ethics for Sign Language Studies” and the related special issue was published in 2009 (Ichikawa et al., 2009). In this symposium, the hearing and the deaf researchers including ordinary people are discussed what is required of the members when they are going to start and publish their researches. Based on the lessons and reflections from the hearing researchers-led sign language studies, new approaches including development of code of ethics (S. Tanaka, 2009a) and collaborative researches with the deaf and the hearing researchers are proposed.

In the field of developing CAE for sign language learning, more than the difference between the professional researchers and the ordinary people, more divers kind of people are involved; the deaf people who use sign language in daily life, the researchers who study sign language or deaf culture, the sign language learners, the children of deaf adults, the sign language teachers, the interpreters and so on. Therefore, in the development for the new version of CASLL, we need to clarify our philosophy and operation policy in easy-to-understand language, and also, we have to realize a system which has a function to protect the teachers and learners from possible violation of the policy, so that all the above people can participate the sign language learning at ease.

In Japan, beside of the ethical issues described above, there is a big argument whether the one should take excessive personal profit or not by teaching JSL as well as the other foreign language teaching businesses. Some people think that teaching JSL should directly connect to training the JSL interpreters and then contribute to better welfare of the deaf people. Therefore, they think the community of JSL teachers should always be recognized by and somewhat under the control of the deaf community. On the other hand, some people think that JSL teachers should be independent from the deaf community and completely free to get more JSL learners.

In the history of JSL, a hierarchical structure with the deaf association on the top and each interpreter training classes on the bottom has been formed to strengthen power of the community. Formulation of the structure made possible to let authorize the JSL interpreter license as one of the national licenses. Therefore, deaf associations cannot admit promotion of the personal JSL teaching business out of their control.

Because of this background, for researchers who have less connection and no authority of the deaf community, it has been difficult to utilize their CAE system into the field. Most of signers who helped researchers to make education materials also resist if the movies with their faces are used for business of their personal profit. For this problem, we always have to care how we can gain trust to our CAE system from the deaf community and how flexibly we can control the training material under the control of both us and signers.

2.3. The Goal

Based the above ethical considerations for building CAE system for sign language learning, we set two goals to develop the new version of CASLL.

1. The signers can rely on the system and willing to be shown in the movies for course materials.
2. Anyone who is interested in JSL learning and also anyone who is interested in teaching JSL from the deaf community can collaborate each other to describe the background knowledge of the JSL.

To reach this goal, we implement following functions to the new CASLL system:

Function to protect the right of publicity In order to prevent the movie materials under control, a function to prevent downloading the movie is needed. There are other options to describe “Creative Commons” or “All Rights Reserved” on the movie to prohibit business use or make the movie under control, but we think that more fundamental design is needed to make the system reliable for all signers who have helped making course materials.

Function to enable collaborative course editing For making CASLL widely used, we think that the material to teach learner about the background of JSL is essential. Since most of existing interpreters’ training class begins their course with teaching deaf culture in advance of the actual teaching of the signs. To create the teaching material which is easy-to-understand for all the learners with different backgrounds, we think a function for collaborative editing is effective. By using this function, people involved in JSL research, movement, and other activities can get together to promote the diffusion of knowledge for JSL beyond each different positions. For JSL teachers, a function for easy editing or adding course materials is also needed.

Function to enhance public relations In order to invite people outside JSL communities, a function for promotion is also needed.

From following sections, we describe the actual implementation of the system.

3. Implementation

3.1. System Architecture

The new CASLL system is implemented by extending open-source Course Management System (CMS) “Moodle”. Moodle has following characters:

- An open-sourced CMS available free of charge.
- Has easy to use course editing interface built in for basic question types.
- Has flexible user management function which can give authority to create a course to specified user.
- Has modular mechanism to extend a function, and has world wide communities where developers are getting together and posting their custom code to extended the functions.
- The design and development of Moodle is guided by a social constructivism pedagogy that enhances a student’s activity in the learning environment (Docs, 2006).

Moodle is the most popular open-source CMS in the world. Also in recent years in Japan, the number of universities are start operating their original e-learning website under the Moodle environment.

However, Moodle does *not* have following functions:

- Contents protection function.
- Question types specific to sign language learning.

As we have discussed in the previous section, above functions are essential to CASLL. To give Moodle the above functionality, we have extended the Moodle by using the module mechanism.

3.2. Contents Protection Extension

Protection of the online contents can be accomplished by several ways.

The first approach is to use a video streaming server which only allows the video player which can communicate to the server using a specific protocol. We can protect the video from downloading, because the user cannot download the video unless using this specific protocol. This approach can protect the video strongly. However, most of the video streaming server is expensive in both computing resources and service price.

The second approach is to use cookie mechanism on the browser. “Cookie” is the default function the browser which can store the information given from the web server. Moodle sets unique string to the cookie when the user logged to the system. We can prevent downloading of the content by only allowing the access from the browser which has specific string in the cookie. However, if the user login to the system, we can not prevent downloading of the content (the user can download the content by right clicking the link on the browser: this is the default content protection policy of Moodle).

The third approach is to use a token with time expiry. We can prevent the content from downloading by generating an

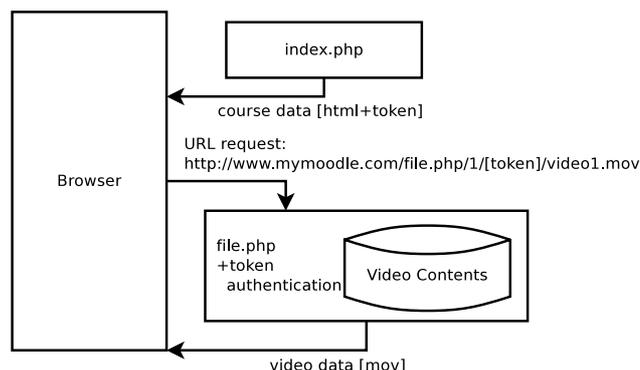


Figure 1: Content protection with token based authentication.

unique URL each time we gives access to the content. By generating an unique URL, the user cannot access to the content by using the same URL as before. This way we can prevent the downloading of the content. This approach generates the unique token with time expiry, and generate the URL based on the token. We only issue the token to the user with specific cookie and who will access the content for the first time. This approach can realize affordable strength content protection in low cost.

We take the third approach. Figure 1 shows the content management mechanism of Moodle. In Moodle, downloading of the contents is done always by calling a single script named “file.php”. Our contents protection extension is realized by applying a patch to the script “file.php”. This patch implements token based authentication function in addition to standard cookie based authentication function in Moodle.

3.3. Sign Learning Question Types Extension

We have implemented sign learning specific question types developed in our previous study(Tanaka et al., 2007a).

User interface for slider and reordering question types are implemented using jQuery UI library¹.

In previous study we only had interface for learners, but in this study we also have implemented an interface for teachers which is integrated to course editing interface of Moodle.

3.4. Wiki Function

The wiki function is originally available as one of Moodle’s default functions. We used this function so that any people can describe and edit the background knowledge which will be needed for the learners working on course materials.

3.5. Link Function

We made a link to the CASLL system from the external multilingualized (sign language enabled) website. Because not only the sign language signer but also the people who are not familiar with sign language will visit the page, multilingualized website is one of the most useful place to invite ordinary people to learn sign language. By showing the logo of CASLL beside the sign language video on the

¹<http://jqueryui.com/>

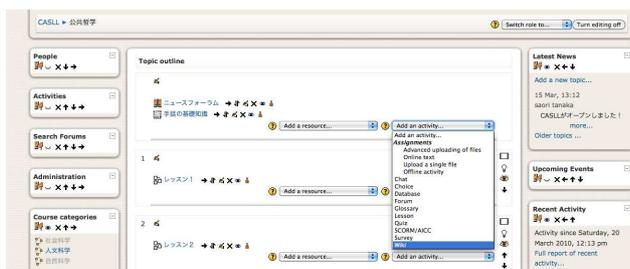


Figure 2: Edit mode of Moodle

website, we have succeed to build a system which attract the visitors' attention into sign language learning in natural way.

4. Current Operation Status

Figure 2 shows a editable page for adding wiki function to CASLL on Moodle. The teachers having the login username and password can enter this page and add the course materials. Each teacher can establish his/her own wiki and also edit the existing wiki to describe about his/her knowledge related to each course material. There are other useful functions. For example, if a teacher wants to ask question to administrator or other teachers, he/she can use forum function in the activity list.

The learning page also have a link from a multilingual website where the original JSL movies are embedded. We are going to increase the same type of link and invite new JSL learners outside existing communities.

5. Conclusion

In this study, we sorted out the problems for the development of online version of CASLL and outlined some ethical issues behind the problems. The difficulties faced to us were one to expose the movies to public, one for increasing the course materials, and one to enhance the collaboration between learners. The ethical discussions revealed that the reliability for the system and the collaborative work for expand the number of course materials were necessary for overcoming these difficulties.

In order to realize the reliable and the collaborative e-learning system, we implemented CASLL within Moodle. We added new functions to Moodle; the protection function for the right of publicity, the wiki function to enable collaborative course editing and finally the Link function to enhance public relations.

Although our development for the online version of CASLL has just started, we are going to evaluate the system design from the view point of the usability for teaching, the effectivity for learning and the utility for collaboration.

6. References

Moodle Docs. 2006. <http://docs.moodle.org/en/Philosophy>.

E. Efthimiou, G. Sapountzaki, K. Karpouzis, and S-E. Fotinea, 2004. *Developing an e-Learning platform for the Greek Sign Language*, pages 1107–1113. Springer, 3118 edition.

Compass: Video for JSL Learning. <http://www.inform-inc.co.jp/video.htm>.

- F. Fukuda. 2005. Compilations of the electronic dictionary of Japanese Sign Language (a second edition) and its instruction manual (in Japanese). *IEICE technical report. Welfare Information technology*, (66):pp.39–44.
- A. Ichikawa, N. Kamei, and K. Kikuchi, editors. 2009. *Japanese Journal of Sign Linguistics*, volume 18. The Japanese Association of Sign Linguistics.
- H. Isono, Y. Takiguchi, M. Katsumata, and M. Nakama. 2006. Preferred reproduction speed of sign language image and audiovisual system via the internet (in Japanese). pages pp.5–8.
- R. Lehman and S. Conceicao. 2001. Involving the deaf community in distance learning, using blended technologies and learning objects. *IEEE Electronic Journal*, pages pp.3–4.
- K. Nakazono and S. Tanaka. 2008. Study of spatial configurations of equipment for online sign interpretation service. *IEICE - Transactions on Information and Systems archive*, (6).
- S. Tanaka. 2009a. From dialogue to the code of ethics (in Japanese). *Japanese Journal of Sign Linguistics*, 18:25–30.
- S. Tanaka. 2009b. The meaning for code of ethics in the interdisciplinary academic societies. In *4th International Conference on Applied Ethics*.
- K. Tabata, T. Kurota, M. Murakami, Y. Manabe, and K. Chihara. 2001. Prototype design for sign language education system (in Japanese). In *Proceedings of Japanese Association of Sign Linguistics*, number 27, pages pp.34–35.
- S. Tanaka, Y. Matsusaka, and K. Uehara. 2007a. Segmentation learning method as a proposal for sign language e-learning (in Japanese). *Human Interface*, 9(2):61–70.
- S. Tanaka, K. Nakazono, M. Nishida, Y. Horiuchi, and A. Ichikawa. 2007b. Skill-nms for an indicator of qualitative skill in the interpreters of japanese sign language. *Proceedings of International Symposium on Skill Science (ISSS)*, pages 178–180.
- S. Tanaka, K. Nakazono, M. Nishida, Y. Horiuchi, and A. Ichikawa. 2008. Evaluating interpreters' skill by measurement of prosody recognition. *Transactions of the Japanese Society for Artificial Intelligence*, 23(3):117–126.

You Get Out What You Put In: The Beginnings of Phonetic and Phonological Coding in the Signs of Ireland Digital Corpus

Gudny Bjork Thorvaldsdottir

Institute of Technology Blanchardstown

Blanchardstown Road North, Dublin 15

thorvalg@tcd.ie

Abstract

The following poster discusses a range of issues with respect to expanding the annotation of the Signs of Ireland (SOI) corpus to incorporate phonetic and phonological coding. This is part of ongoing PHD research that explores the phonology-morphology interface in ISL. It is the intention to identify the phonemes and the allophones of ISL using the corpus and thus it is necessary to incorporate a detailed annotation at the phonetic level. To date, no research has been done in this area apart from a phonetic description of handshapes in the language. The poster outlines how a range of phonetic features have been established for ISL, drawing on work on other signed languages, and the changes that had to be made to the original list of features to accommodate ISL. Also discussed are the factors influencing decisions regarding the coding and naming of handshapes at phonetic level and what type of tiers were needed to accommodate the proposed research and future research at the phonetic and phonological level.

1. Introduction

This poster discusses a range of issues with respect to expanding the annotation of the Signs of Ireland (SOI) corpus to incorporate phonetic and phonological coding. This forms part of ongoing PHD research work that explores the phonology-morphology interface in Irish Sign Language (ISL).

The SOI corpus consists of over 40 narratives that have already been highly annotated: it contains glossed lexical signs, classifier constructions and non-manual features. Classifier handshapes have also been annotated. It is my intention to identify the phonemes and the allophones of ISL using the corpus and it is thus necessary to incorporate a detailed annotation at the phonetic level.

This poster outlines how, by drawing on Crasborn's (2001) and Van der Kooij's (2002) work on Sign Language of the Netherlands (SLN), a list of phonetic features have been established for ISL and the changes to the original list of features that were required in order to accommodate ISL.

I also outline the factors influencing decisions regarding the coding and naming of handshapes at phonetic level. These include the question of whether already established naming conventions be maintained. For example, moving away from established protocols will result in inconsistencies within the annotations in the corpus. However, for the purposes of phonetic research a more elaborate coding might be necessary. Another challenge involves establishing what types of tiers are needed to accommodate the proposed research as well as future research at the phonetic and phonological level.

2. Phonetic Features for ISL

In order to identify the phonemes and the allophones of ISL, a list of phonetic features for the language must be identified. To date, no research has been done in this area apart from basic work describing handshapes in ISL. Thus far, there is no agreement on the phonetic alphabet inventory for ISL: Ó'Baoill and Matthews (2000) identified 66 handshapes while Matthews (2005) identified 78. The issue of allophonic variation has not yet been tackled for this language.

The other parameters that have traditionally been used to describe signs (i.e. location, movement and orientation) have not been researched in ISL at phonological or morphological level. All that currently exists is a vaguely phonetic level description of parameters with respect to research on American Sign Language (ASL) (See O’Baioill and Matthews, 2000; Matthews, 2005).

Since there is no detailed list of phonetic features in ISL existing, we will incorporate work that has been done on SLN (Crasborn, 2001; Van der Kooij, 2002) and ASL (e.g. Stokoe, 1960; Liddell and Johnson, 1989). By drawing on this work we have established a list of phonetic features for ISL. Because we do not have a precise knowledge of what phonetic features exist in ISL, apart from handshapes, and we do not yet know which properties may be distinctive in the language, we have initially included a vast array of phonetic properties. As the work proceeds then, we expect this list to be reduced.

2.1 ISL handshapes

For annotation purposes, challenges arise in terms of how handshapes are recorded: for example, of the 66 handshapes identified in Ó’Baioill and Matthews (2000), 28 are established as occurring as classifier handshapes also. These are annotated following ECHO project annotation norms (Nonhebel et al., 2004) where possible, with additional handshapes drawn from a list of 48 classifier handshapes described for BSL in Brennan (1992) using names like CL-B, CL-ISL-K etc. within the framework of the SOI corpus.

There is some inconsistency in the literature when it comes to handshape names. Researchers usually use names that refer to the alphabet in the sign language being discussed. Although some of these names are compatible between signed languages, such as B (a flat hand) and A (a fist-handshape), we do find different naming conventions as well (e.g. W in SLN uses thumb, index and middle finger which is represented as 3 in ASL). For transcription purposes, we have decided to incorporate the coding used in the SignPhon database¹ (A1, A2, B1, B2 etc., see van der Kooij 2002). This will save time when transcribing and is useful if we later decide to use SignPhon to create a database for lexical signs in ISL. Also, coming up with names for all ISL handshapes is a time consuming process and redundant at this stage since we expect this list to change as the research proceeds.

¹ This is a database created to research phonetics and phonology of SLN and includes lexical signs only (See Crasborn 2001; Crasborn et al. 2001; van der Kooij 2002).

Some changes have already been made to our current list of handshapes (see figure 1 a-b).

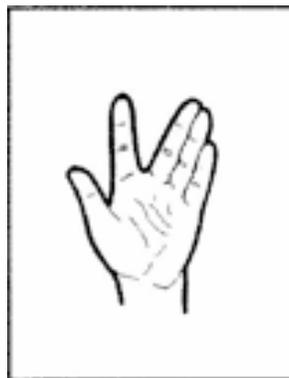


Figure 1a²: ISL handshape not found in HamNoSys (Prillwitz et al., 1989)



Figure 1b³: Handshape not noted before in ISL, but used in Signs like BOY.

Thus, the naming conventions for classifier handshapes in the corpus have not been maintained for lexical signs. In order to facilitate search between handshapes in lexical signs and CCs⁴, information on the names of classifier handshapes is included in the notes tier. A subdatabase for handshapes, drawing on SignPhon, will be created where the exact articulation of the handshape and semantic information is included.

3. Discussion

As noted above, one challenge involves establishing what types of tiers are needed for the research. When attempting to transcribe or code phonetic features in a language with the aim of using the information in phonological analysis, a problem

² Illustration copyright © Patrick Matthews (forthcoming).

³ Handshape figure from Prillwitz et al. (1989)

⁴ Classifier Constructions.

irises as how to make the coding functional when doing different searches regarding phonology. Ideally then, one should know the phonology of the language and what kind of search will be necessary *before* attempting the phonetic coding. However, this is seldom the case. This problem has been referred to as the *database paradox* by Crasborn et al. (2001) and Van der Kooij (2002). In order to beat this paradox, it is necessary to rely on research in other signed languages as well as preliminary observation of the language in question.

Crasborn et al. (2001) report that a disadvantage of the SignPhon database is that it includes one instance of a sign, articulated by one signer, thereby excluding the possibility of variation being detected: “ideally, to make a phonological analysis one would want to compare different instances of the same sign, signed by various signers in various contexts” (p 224). While this is possible in the SOI corpus, it poses another problem which relates to the kind of data we are using for this research. The SOI corpus is a ‘live’ corpus and therefore the signs are not articulated in citation form. Thus, we must ask how variation in articulation can be annotated in the corpus so that they are still identifiable in a search. Figure two a-b shows an example of variation in SF⁵ articulated by the same signer within one narrative. Interestingly, the sign articulated *before* the variant of BOY in figure two b, is a two handed sign using a handshake with four SF (the remnants of the sign can still be seen on the non-dominant hand), thus ruling out an instance of assimilation. In order to detect such instances of variation in a search, we have included a tier for phonetic variation where the ‘correct’ feature is noted.



Figure 2 b: Variation of the sign BOY, articulated with one SF.



Figure 2a: The sign BOY, articulated with four SF.

⁵ SF = Selected Fingers

4. References

- Brennan, M. (1992). The Visual World of British Sign Language: An Introduction. In: Brien, D. (ed.) *Dictionary of British Sign Language/English*. London: Faber & Faber.
- Crasborn, Onno. (2001). *Phonetic Implementation of Phonological Categories in Sign Language of the Netherlands*. Utrecht: LOT. PhD Dissertation.
- Crasborn, O., Hulst, van der, H. and Els van der Kooij. (2001). SignPhon: A phonological database for sign languages. in: *Sign Language and Linguistics* 4:3/2, 215-28. John Benjamins Publishing Company.
- Kooij, E. van der. (2002). *Phonological categories in Sign Language of the Netherlands: phonetic implementation and iconic motivation*. Utrecht: LOT. PhD Dissertation.
- Liddell, Scott & R. Johnson. (1989). American Sign Language: The phonological base. *Sign Language Studies* 64. 197-277.
- Matthews, P.A. (2005). Practical Phonology: What learners need to know about handshapes in Irish Sign Language. *Deaf Worlds, International Journal of Deaf Studies, Volume 21, issue 2*. Douglas McLean.
- Matthews, P. A. (forthcoming). ISL Communication – Unit 1 and Unit 2. Dublin.
- Nonhebel, A., Crasborn, O. & E. van der Kooij. (2004). Sign language transcription conventions for the ECHO project. ECHO project, University of Nijmegen. URL: http://www.let.kun.nl/sign-lang/echo/docs/transcr_conv.pdf
- O’ Baoill, Donall and Matthews, P.A. (2000). *The Irish Deaf Community, Volume 2: The Structure of Irish Sign Language*. Dublin: ITE.
- Prillwitz, S., R. Leven, H. Zienert, T. Hanke, J. Henning et al. (1989). *HamNoSys version 2.0. Hamburg notation system for sign languages: an introductory guide*. Hamburg: Signum.
- Stokoe, W. C. (1960). Sign language structure: and outline of the visual communicatin systems of the American deaf. In: *Studies in Linguistics: Occasional Papers*. Buffalo: University of Buffalo.

ATLAS Project: Forecast in Italian Sign Language and Annotation of Corpora

Mara Vendrame¹, Gabriele Tiotto²

(1) Università degli Studi di Torino, Dipartimento di Psicologia
Via Po 14, 10123 Torino, Italy

E-mail: mara.vendrame@unito.it

(2) Politecnico di Torino, Dipartimento di Automatica e Informatica
Corso Duca degli Abruzzi 24, 10129 Torino, Italy

E-mail: gabriele.tiotto@polito.it

Abstract

The paper presents the preliminary results of a research project focused on the creation and the annotation of one Italian Sign Language corpus concerning the weather forecasts domain. As a result of the annotation process, our annotations of signs sequences showed that the semantics of the signed discourse cannot be grasped just through an annotation of single weather signs which exploits the five parameters *handshape*, *movements*, *directions*, *locations* and *non-manual components*. Rather, from the annotation process appears that, in order to grasp the discourse semantics, it is necessary to consider the extensive use of Highly Iconic Structures in order to specify the iconic properties of the different atmospheric phenomena. In particular, it often occurs that several signs are combined among themselves (see also Cuxac, 2000; Di Renzo, et al, 2006; Pizzuto et al., 2008; Pizzuto, Rossini & Russo, 2006). Thus, respect to single signs, our analysis of complex manual and non-manual units stored in our database suggests the necessity to better explore multidimensional aspects, in order to properly develop and train an automatic translator able to translate from Italian written text to Italian Sign Language.

1 Introduction - ATLAS Project: purposes and characteristics

Our study is part of the Automatic Translation into Sign Languages (ATLAS) Project, targeting the development of several tools to provide signing deaf people full access to broadcast communications. In order to include and let signing deaf people to proactively collaborate in the global community, this project will grant a wide range of services such as the possibility to follow and understand media information delivered in Italian Sign Language.

As the cost of translation services furnished by a human interpreter is very high, the reason for creating an automatic translation system is the economic advantage. In particular, ATLAS focuses on the creation of an automatic translator from written Italian texts to Italian Sign Language through an intermediate translation in a written form of the Italian Sign Language.

Nowadays no Italian Sign Language weather forecast service exists, our study aims at making good of this deficit, to allow signing deaf individuals to access to weather forecast news in their mother tongue.

2 Signing deaf individuals' difficulties with spoken and written verbal languages

Sign language is the visual-spatial language of signing deaf individuals (Emmorey, 2002). Through sign languages deaf individuals become members of the Deaf community which are widespread all over the world. As members of the Deaf community, deaf individuals consider their sign language a crucial aspect of their cultural identity (Padden and Humphries, 1988).

Signing deaf individuals have no problems in understanding their mother tongue (Pizzuto, Caselli & Volterra, 2000; Sacks, 1990). On the contrary, all verbal

languages are difficult for deaf individuals to understand. The literature reveals that signing deaf individuals have difficulties with spoken and written language, and this claim holds also for signing deaf Italian individuals (Arfè, 2003; Fabbretti & Tomasuolo, 2006; Pizzuto et al., 2000). Indeed, sign languages differ from spoken languages on several dimensions. All visual-gestural languages possess a rich morphosyntactic structure organized in space, which differs from the sequential ordering of the sentence elements in verbal languages (Bagnara et al., 2008; Russo Cardona & Volterra, 2007; Volterra, 2004).

In particular, the morphosyntactic elements in sign languages are effectively conveyed through facial expressions, body posture and spatial resources, whereas in verbal languages these elements are conveyed through function words like prepositions, articles, conjunctions. As a result, when reading and processing written texts, signing deaf individuals possess scant ability to process basic grammatical morphemes (such as articles, prepositions, conjunctions, pronouns, and verbal auxiliaries), which lead them to a poor exploitation of the semantic and the pragmatic information necessary to reconstruct the meaning of the global message (Radelli, 1998; Vendrame, Cutica & Bucciarelli, 2009; Volterra, Capirci & Caselli, 2001).

3 The weather domain: the creations of news signs

We started to analyze the Italian version of fifty original written weather texts provided by RAI Italian national television. We pointed out some of their peculiar characteristics, such as a formal language with complex sentence structures, the high presence of technical weather related words and frequent references to cardinal points.

The fifty texts were translated into Italian Sign

Language by a sign language interpreter. In particular, the interpreter translated the written Italian texts into the national Italian Sign Language as defined in Radutzky's (2001) dictionary. Thus, for example, we adopted the weather standard signs contained in the Radutzky dictionary for *sun*, *wind*, *snow* and *rain*.

As the Italian Sign Language has no specific signs to describe the atmospheric, a team composed by one hearing interpreter with a group of native deaf signers created a list of new weather signs for those atmospheric events which have not a corresponding sign in the Italian Sign Language. Further, for some standard Radutzky's signs the team created several graduated signs.

For example, the standard Radutzky's sign *rain* was modified in order to express both *misty rain*, *downpour*, and *storm*. The comprehensibility of such new signs was ascertained with other interpreters from different Italian regions. The interpreter was video-recorded while signing each weather sign in a neutral space. As a final step, the interpreter was video-recorded while signing each weather news forecast.

Finally we analyzed the videos of five weather forecasts: our manual and software aided annotation focused on the combination of the five parameters *handshape*, *movements*, *directions*, *locations* and *non-manual components*.

4 Annotation difficulties

Our annotation task posed many problems, due to the fact that respect to verbal languages annotations, sign languages annotations involve a meta-linguistic task in order to grasp the multidimensional aspects of sign languages (Pizzuto et al., 2008). First of all, respect to our previous annotation of single individual signs, annotation of sentences became rapidly a difficult task. We had to decide what exactly is relevant for producing an accurate annotation, and what we could leave aside. In particular, which aspects of manual and non-manual features had to be considered in order to implement an automatic translator from written Italian language to Italian Sign Language? Indeed, grammatical information in Italian Sign Language are clearly conveyed through spatial modifications of the same sign.

In line with Di Renzo and colleagues (2006), our main difficulty was to describe streams of signs tightly linked to each other as in sign language discourse. In particular, due to co-articulation phenomena, we noted that the beginning of a sign is modified according to the previous sign, and the end of the same sign is modified according to the following sign (see also Pizzuto, 2003; Segouat, 2009).

The signed units annotation revealed two main structural features of the visual-spatial lexicon and grammar of Italian Sign Language for the weather: a high presence of re-locable signs due to spatial cardinality, and interrelated compound signs.

4.1 Multidimensional representations of weather scene

In the Italian weather texts, cardinal points and spatial references are described in a linear manner, whereas in the parallel versions of the Italian Sign Language, they are expressed simultaneously and multidimensionally.

In line with other studies (Pizzuto et al., 2008;

Cuxac, 2000), our annotation had to grasp structural features, unique to the sign languages (Pizzuto, 2007; Pizzuto & Pietrandrea, 2001) and represented through manual and non-manual elements arranged in a multidimensional and in multilinear fashion.

Consider, for example, the following sentences: "Local and light cloudiness could take place in the north-eastern sector, then starting from the evening, an increase in cloudiness on the western one".

Cardinal points in sentences were not represented by standard elements, such as through the index finger directed towards the cardinal points, but through complex signs structures dislocated in space with body shift and eye gaze directions towards left or right, up or down. The interconnections of these elements was able to communicate "the whole weather situation" in a simultaneous manner.

4.2 Iconic structures

We found an high presence of non-standard constructions, namely a high presence of highly iconic structures with manual and non-manual features devoted to reproduce the embodied entity (Cuxac, 2000).

In particular, in weather domain, we noted two types of transfer: transfer of form and size, and transfer of situation. Both types are common in signed discourse, in signed poetry and in signed narratives (Pizzuto, 2007; Russo, Giuranna & Pizzuto, 2001). Transfer of form describes objects or persons according to their size or form, transfer of situation involves the shift of a sign referring to either an object or a character relative from a stable locative point of reference (Cuxac, 2000; Pizzuto et al., 2008; Sallandre, 2003). In signed sentences, the presence of iconicity has a crucial role, because it allows the interpreter to describe in a comprehensible way the atmospheric events according to their size or form. For example, in order to communicate salient differences between "nebula" and "clouds lied around", the interpreter does not use standard signs, but adopts "productive" highly iconic constructions, which describe in a iconic manner the different forms of the clouds.

Consider, for example, the following sentence contained in one text: "Today in southern regions we saw thunderstorms, which gradually weakened, some improvements in the Adriatic area". In order to describe the weather situation, the interpreter utilizes a transfer of situation structure, in which manual and non manual units can be combined among themselves, and they result in a dynamic depiction of the weather situation. Further, the use of situation transfer is accompanied by specific eye-gaze pattern which are oriented towards the hands, and by specific facial expressions (Pizzuto et al, 2008). More in general the weather situation exists as it was observed from a distance (Pizzuto, 2007).

We noted a multilinear organization of information whereby two referents can be simultaneously specified, and also maintained in time and space in a modality that appears to be unique of sign language (Pizzuto et al., 2008). Further, the situation transfer is accompanied by locative point of reference.

As the weather bulletin texts are characterized by geographical coordinate, we remarked an high presence of two manual indexes in order to provide references to cardinal points, accompanied by a gaze pointing in the

same direction.

Thus, in line with previous studies (Di Renzo et al., 2006) we first outline how our annotation have to describe complex sign units that are very frequent in sign languages discourse, and exhibit highly iconic and multilinear features, that have no reference in verbal languages (Pizzuto et al., 2008).

Consider, for example, the following sentences contained in the text: "Ionian sea is very heavy, generally heavy the other seas, bit heavy only the basins to north". In this case the interpreter's translation is characterized by a transfer of situation: some manual and non-manual components are simultaneously arranged in time and space to represent the shift "from heavy seas to a bit heavy seas" (Cuxac, 2000). Further, as we noted previously, non-manual components such as cheek's blow up, left half open eyes and half-mouth are congruent with the process represented.

Analyzing these elements we had the possibility to detect different typologies of signing "styles". They can be classified as:

- Signed Italian
- Polluted Italian Sign Language
- Pure Italian Sign Language

These three typologies are detected and classified with respect to the amount of iconic, incorporation and multidimensional elements in the signing act.

Signed Italian is poor of iconic structures and the use of multidimensional representation is limited. In this case use of the facial expressions and incorporation is limited. "Polluted Italian Sign Language" can be seen as a signed Italian in which there is a frequent use of iconicity and multidimensionality but is in some way polluted by elements proper of Signed Italian. Facial expression is used but we detected a low use of incorporation. Pure Italian Sign Language is the preferred communication modality of deaf people and is rich in iconicity, incorporation and for this reason is extremely dynamic (i.e. a single sign can be signed in different ways).

These considerations make relevant to choose the right tradeoff between quality of the representation and complexity in annotation. In order to provide the best translation possible, we decided to create and annotate the movies in Pure Italian Sign Language.

This in line with the ATLAS project objectives that tries to provide a complete translation resorting to the Italian Sign Language grammar.

5 The annotation of video content in the weather forecast domain

A study on previous project targeting sign language annotation had been performed in order to derive guidelines for the annotation of our weather forecast content.

The automatic translation purpose makes relevant to provide the statistical translator all the needed information for the parameterization of the signs. Since they present modification within utterances with respect of their basic lexical form these information have to be notated. Iconicity, co-articulation and the relationship between the signed entities are part of the semantics of the signed discourse and have to be described during annotation.

After several studies we created a formalism that can be considered an annotation schema. We have not to neglect that this formalism conveys also visualization information that can be provided to the system modules devoted to convert linguistic content to character animation movements.

A detailed description of this formalism is out of the scope of this paper but it worth to point out that the advantages of applying this formalism to annotation are that the annotator is in some way guided to annotate just the necessary information for the automatic translation and for a complete description of the signs. On the other side it is rich enough to provide the basis for the development of a complete knowledgebase.

The annotation is performed using a custom built annotation tool that is based on our formalism. This is able to store the information in a database that includes the Radutzky Italian Sign Language dictionary, the ATLAS dictionary with signs within the weather forecast domains and other non standard signs. This provides a knowledgebase for the creation of the Italian Sign Language corpus.

6 Conclusions

Even if our study is still ongoing, our annotation revealed that, as in face to face sign language modality, also in weather domain high spatial arrangement, facial expressions and iconic structures, are the most peculiar components.

Thus, with respect to standard signs listed in Radutzky Italian Sign Language dictionary, and isolated new weather signs, our annotation have to properly consider complex sign constructions with complex meaning that are very frequent in signed discourse, and grammar as part of the non-standard or productive lexicon (Cuxac, 2000; Di Renzo et al., 2006; Pizzuto et al., 2008). The attempt to create the first Italian Sign Language corpus in Italia made relevant the considerations pointed by previous studies.

The creation of new signs required the definition of a roadmap in order to consider the linguistic and cognitive issues, in a non standard domain in Italian Sign Language. The roadmap to the creation affected also the procedures for annotation, since new issues enriched the formalism that supports the representation of a written form of Italian Sign Language and the development of the annotation tool.

7 Acknowledgements

The work presented in the present paper has been developed within the ATLAS (Automatic Translation into sign LAnAgeS) Project (ID 44), co-funded by Regione Piemonte within the "Converging Technologies - CIPE 2007" framework (Research Sector: Cognitive Science and ICT).

8 References

- Arfè, B. (2003). La produzione del testo in persone sorde: aspetti linguistici e cognitivi del processo di scrittura. *Psicologia Clinica dello Sviluppo*, 1, pp. 7-28.
- Bagnara, C., Corazza, S., Fontana, S., & Zucalà, A. (2008). *I segni parlano. Prospettive di ricerca sulla Lingua dei Segni Italiana*. Milano: Franco Angeli.
- Cuxac, C. (2000). *La Langue des Signes Francaise (LSF)*.

- Les voies de l'iconocité*. Faits de Langues, n. 15-16. Paris: Ophrys.
- Di Renzo, A., Lamano, L., Lucioli, T., Pennacchi, B., & Ponzo, L. (2006). Italian Sign Language: Can we write it and transcribe it with Sign Writing? in C. Vettori (ed.), *Proceedings of the Second Workshop on the Representation and Processing of Sign Languages, International Conference on Language Resources and Evaluation*, LREC, Genoa, May 28th 2006. Pisa: ILC-CNR, pp. 11-16.
- Emmorey, K. (2002). *Language, cognition, and the brain: Insight from sign language research*. Mahwah, N.Y: Erlbaum.
- Fabbretti, D., & Tomasuolo, E. (2006). *Scrittura e sordità*. Roma: Carocci.
- Padden, C., & Humphries, T. (1988). *Deaf in America: Voices from a Culture*. Cambridge: Harvard University Press.
- Pizzuto, E. (2003). Coarticolazione e multimodalità nelle lingue dei segni: dati e prospettive di ricerca dallo studio della Lingua dei Segni Italiana (LIS). In Marotta G., Gnocchi N. (a cura di), *La coarticolazione - Atti delle XIII Giornate GFS*, Pisa, Edizioni ETS, pp. 59-77.
- Pizzuto, E. (2007). Deixis, anaphora and person reference in signed languages. In E. Pizzuto, P. Pietrandrea & R. Simone (eds.), *Verbal and Signed Languages - Comparing structures, constructs and methodologies*. Berlin/ New York: Mouton De Gruyter, pp. 275-308.
- Pizzuto, E., Caselli, M.C., & Volterra, V. (2000). Language, Cognition and Deafness. *Proceedings from Newborn Hearing Systems, Seminars in Hearing*, 21, pp. 343-358.
- Pizzuto, E., Pietrandrea, P. (2001). The notation of signed texts: open questions and indications for further research. *Sign Language and Linguistics, (Special Issue on Sign Transcription and Database Storage of Sign Information)*, 4: 1/2, pp. 29-43.
- Pizzuto, E., Rossini, P., & Russo, T. (2006). Representing signed languages in written form: questions that need to be posed. In C. Vettori (ed.), *Proceedings of the "Second Workshop on the Representation and Processing of Sign Languages" - LREC 2006 - 5th International Conference on Language Resources and Evaluation*. Genoa, May 28th 2006, Paris: ELRA2006, pp. 1-6.
- Pizzuto, E., Rossini, P., Sallandre, M.A., & Wilkinson E. (2008). Deixis, anaphora and Highly Iconic Structures: Cross-linguistic evidence on American (ASL), French (LSF) and Italian (LIS) Signed Languages. In R. Müller de Quadros (Ed.), *Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th Theoretical Issues in Sign Language Research Conference* (pp. 475-495). Petrópolis/RJ, Brazil: Editora Arara Azul.
- Radelli, B. (1998). Nicola vuole le virgole. Introduzione alla logogenia. Padova, Decibel.
- Radutzky, E. (a cura di) (2001). *Dizionario bilingue elementare della Lingua Italiana dei Segni*. Edizioni Kappa, Roma.
- Russo Cardona, T., & Volterra, V. (2007). *Le lingue dei segni. Storia e semiotica*. Roma, Carocci Editore.
- Russo, T., Giuranna, R., & Pizzuto, E. (2001). Italian Sign Language (LIS) poetry: iconic properties and structural regularities. *Sign Languages Studies*, Vol.2., 1, Fall 2001, pp. 84-112.
- Sacks, O. (1990). *Vedere voci. Un viaggio nel mondo dei sordi*. Milano, Adelphi.
- Sallandre, M-A. (2003). Les unités du discours en Langue des Signes Française. Tentative de categorization dans le cadre d'une grammaire de l'iconicité. Ph.D. Dissertation, Paris, Université Paris 8.
- Segouat, J. (2009). A study of Sign Language coarticulation. *Sigaccess Newsletter*, 93, pp. 31-38.
- Stokoe, W. (1960). Sign language structure: An outline of the visual communication systems of the American deaf. *Studies in Linguistics, Occasional Paper* (8) (2nd printing 1993, Burtonsville, MD: Linstok Press).
- Vendrame, M., Cutica, I., & Bucciarelli, M. (2009). Learning by models from texts and from videos: where deafs differ from hearings. *Proceedings of the British Psychological Society, Annual Conference, Cognitive Psychology Section*. Hatfield, England, 1-3 September 2009, p. 71.
- Volterra, V. (a cura di) (2004). *La lingua italiana dei segni*. Bologna: Il Mulino.
- Volterra, V., Capirci, O., & Caselli, C. (2001). What atypical populations can reveal about language development: The contrast between deafness and Williams syndrome. *Language and Cognitive Processes*, 16, pp. 219-239.

SIGNUM Database: Video Corpus for Signer-Independent Continuous Sign Language Recognition

Ulrich von Agris and Karl-Friedrich Kraiss

Institute of Man-Machine Interaction, RWTH Aachen University, Germany
{vonagris,kraiss}@mmi.rwth-aachen.de

Abstract

Research in the field of continuous sign language recognition has not yet addressed the problem of interpersonal variance in signing. Applied to signer-independent tasks, current recognition systems show poor performance as their training bases upon corpora with an insufficient number of signers. In contrast to speech recognition, there is actually no benchmark which meets the requirements for signer-independent continuous sign language recognition. Because of this absence we created a new sign language corpus based on a vocabulary of 450 basic signs in German Sign Language (DGS). The corpus comprises 780 sentences each performed by 25 native signers of different sexes and ages. This database is now available for all interested researchers.

1. Introduction

The development of automatic sign language recognition systems has made significant advances in recent years. Research efforts were mainly focused on robust extraction of manual and non-manual features from the signer's articulation. Additional attention was paid to classification methods. First implementations proved that using subunit models has advantages over word models when recognizing large vocabularies.

The present achievements provide the basis for future applications with the objective of supporting the integration of deaf people into the hearing society. Translation systems and automatic indexing of signed videos are just two examples. Further applications arise in the field of human-computer interaction. Multimodal user interfaces and the control of human avatars could be realized via gesture and mimic recognition.

All these applications have in common that they must operate in a user-independent scenario. Current systems for sign language recognition achieve excellent performance for signer-dependent operation. But their recognition rates decrease significantly if the signer's articulation deviates from the training data.

Interpersonal variability The performance drop in case of signer-independent recognition results from the strong interpersonal variability in production of sign languages. Even within the same dialect, considerable variations are commonly present. Figure 1 shows different articulations of an exemplary sign in British Sign Language.

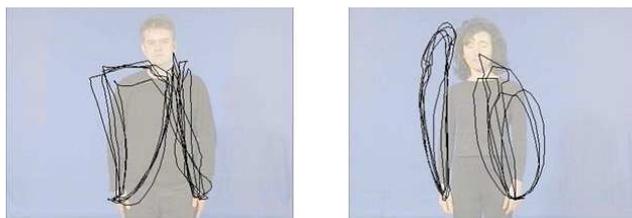


Figure 1: The sign 'tennis' performed five times by two different native signers using the same dialect. Positions of the hands are visualized as motion traces for comparison.

Analysis of the hand motion reveals that variation between different signers is significantly higher than within one signer. Other manual features such as hand shape, posture, and location exhibit analogue variability.

2. The SIGNUM Project

Although signer-independence is an essential precondition for future applications, only little investigations have been made in this field so far. This unexplored gap was subject of a research project called SIGNUM (Signer-Independent Continuous Sign Language Recognition for Large Vocabulary Using Subunit Models), funded by the Deutsche Forschungsgemeinschaft. The project was carried out by the Institute of Man-Machine Interaction, located at the RWTH Aachen University in Germany. It aimed to develop a video-based automatic sign language recognition system that allows signer-independent continuous recognition.

System Overview Following sign language recognition system constitutes the basis for our ongoing research work. A thorough description is given in (Kraiss, 2006; von Agris et al., 2008c). The system utilizes a single video camera for data acquisition to ensure user-friendliness. Since sign languages make use of manual and facial means of expression both channels are employed for recognition.

For mobile operation in uncontrolled environments sophisticated algorithms were developed that robustly extract manual and facial features. The extraction of manual features relies on a multiple hypotheses tracking approach to resolve ambiguities of hand positions (Zieren and Kraiss, 2005). For facial feature extraction an active appearance model is applied to identify areas of interest such as the eyes and mouth region. Afterwards a numerical description of facial expression, head pose, line of sight, and lip outline is computed (Canzler, 2005).

Based on hidden Markov models the classification stage is designed for recognition of isolated signs as well as of continuous sign language. In the latter case a stochastic language model can be utilized, which considers uni- and bigram probabilities. For statistical modeling of reference models each sign is represented either as a whole or as a composition of smaller subunits – similar to phonemes in spoken languages (Bauer, 2003).

As the articulation of a sign is subject to high interpersonal variance dedicated adaptation methods known from speech recognition were implemented and modified to consider the specifics of sign languages. For rapid signer adaptation the recognition system employs a combined approach of eigen-voices, maximum likelihood linear regression, and maximum a posteriori estimation (von Agris et al., 2008a).

3. Related Work

The realization of a signer-independent recognition system requires a database containing training material with articulations of a large number of different signers. The more signers articulate the same signs the better will be the overall recognition performance after training.

The reader interested in a survey of the current state in sign language recognition is directed to (Ong and Ranganath, 2005). Similar to the early days of speech recognition, most researchers focus on the recognition of isolated signs. Only a few recognition systems were reported that can process continuous signing. Here most research was done within the signer-dependent domain, i.e. every user is required to train the system himself before being able to use it. Most sign language corpora solely contain articulations of a single signer and are therefore not suited for training signer-independent systems.

In total only three corpora (Fang et al., 2002; Zahedi et al., 2006) reported in literature comprise sentences articulated by more than one signer. However, these databases are of limited use as they do not sufficiently cover interpersonal variance due to following reasons. In the case of the ASL corpus in (Zahedi et al., 2006) and the CSL corpus in (Fang et al., 2002) the number of signers is by far too small. Moreover both corpora reported in (Zahedi et al., 2006) include a large number of signs that occur only once or twice in the whole dataset. Obviously, these signs were not performed by all signers but only by a maximum of two signers. This results in the same problem that the number of signers is not sufficient for training signer-independent models.

In summary, it can be stated that none of the corpora currently found in literature meets the requirements for signer-independent continuous sign language recognition. In contrast to speech recognition, there is actually no standardized benchmark.

4. The SIGNUM Database

For this reason we decided to create a new sign language corpus, which should be made available for other interested researchers after the project ends. We hope that the release of this database will boost research efforts in the fields of sign language recognition. Maybe it will become established as the first benchmark for signer-independent continuous sign language recognition.

Since we use a vision-based approach for sign language recognition the corpus was recorded on video. Table 1 summarizes the most important details about our corpus.

4.1. Corpus Concept

The SIGNUM Database contains videos of isolated signs and of continuous sentences performed by various signers. The vocabulary comprises 450 signs in German Sign

General Information	
Name:	SIGNUM Database
Author:	Ulrich von Agris
Recording:	2007 - 2008
Production status:	Completed

Corpus Content	
Language:	German Sign Language
Vocabulary size:	450 basic signs
Number of signers:	25 native signers
Number of signs:	450
Number of sentences:	780
Number of performances:	
- Reference signer	3
- Other signers	1
Total number of sequences:	33,210
Equivalent video duration:	55.3h

Technical Details	
Image resolution:	776 × 578, 30fps, color
Image format:	JPEG (8:1 compression)
Data volume:	920GB (approx.)

Resource Availability	
Data centers:	BAS, ELRA
Documentation:	Online

Table 1: Important details about the SIGNUM Database.

Language representing different word types such as nouns, verbs, adjectives, and numbers. Those signs were selected which occur most frequently in everyday conversation and are not dividable into smaller signs. Hence, they are called basic signs in the following. For selection several books and visual media commonly used for learning German Sign Language were evaluated.

All 450 basic signs differ in their manual parameters. Many of them, however, change their specific meaning when the manual performance is recombined with a different facial expression. For example, the signs BÜRO (OFFICE) and SEKRETÄRIN (SECRETARY) are identical with respect to gesturing and can only be distinguished by the signers lip movements. In this case only the former sign is regarded as basic sign, whereas both signs appear in the continuous sentences of the corpus. In total 134 additional signs, derived from the basic signs, were integrated into the corpus. Furthermore, some of the basic signs can be concatenated in order to create a new sign with a different meaning. For example, the sign KOPF+SCHMERZEN (HEADACHE) is composed of the two basic signs KOPF (HEAD) and SCHMERZEN (PAIN). According to this concept, 156 composed signs were collected and integrated as well. Although the selected vocabulary is limited to 450 basic signs, in total 740 different meanings can be expressed by means of recombination and concatenation.

Based on this extended vocabulary, overall 780 sentences were constructed. No intentional pauses are placed between signs within a sentence, but the sentences themselves are separated. Each sentence ranges from two to eleven signs in length. All sentences are grammatically well-formed. The annotation follows the specifications of the Aachener Glossenumschrift, developed by the Deaf Sign Language Research Team (DESIRE) at the RWTH Aachen University (DESIRE, 2004).

In order to evaluate the recognition performance for different vocabulary sizes, the corpus is divided into three sub-corpora simulating a vocabulary of 150, 300, and 450 basic signs respectively.

4.2. Interindividual Variation

For modeling interindividual variation in articulation all 450 basic signs and 780 sentences were performed once by 25 native signers of different sexes and ages. One of them was chosen to be the reference signer. His articulations were recorded even three times, serving for evaluation of the signer-dependent recognition rates. In total 33,210 utterances (12,150 signs and 21,060 sentences) are stored in the database.

Subjects were recruited in the western parts of Germany by placing advertising posters in several institutions visited primarily by deaf people. Each subject read and signed a project consent form. For 80% of the signers German Sign Language is their native language. Almost all of them attended school in Germany and have at least very good sign language skills. Table 2 gives some statistics about their personal data (sex, age, body size, body weight, hearing status, and dominant hand).

Sex	
Male:	12
Female:	13

Age	
21-25 years:	8
26-30 years:	9
31-40 years:	6
41-50 years:	2

Body size	
1.51-1.60 m:	3
1.61-1.70 m:	6
1.71-1.80 m:	10
1.81-1.90 m:	6

Body weight	
51-60 kg:	4
61-70 kg:	6
71-80 kg:	6
81-90 kg:	4
91-99 kg:	1
unknown:	4

Hearing status	
Deaf:	23
Hearing impaired:	2

Dominant hand	
Right:	23
Left:	2

Table 2: Some statistics about the signers' personal data.

4.3. Recording Conditions

In order to facilitate feature extraction video recordings were conducted under laboratory conditions, i.e. controlled environment with diffuse lighting and a unicolored blue background (see Figure 2). The scene was illuminated

frontally by six fluorescent lamps, each equipped with two tubes generating true natural daylight. Diffusion filters were mounted in front of the lamps for spreading the light beam and reducing shadows.



Figure 2: Example frame taken from the reference signer.

The signers wear dark clothes with long sleeves and perform from a standing position. Moreover each signer was instructed to move his hands from a resting position beside the hips to the signing location and after signing back to the same resting position. The hands are visible throughout the whole sequence, and their start and end positions are constant and identical which simplifies tracking.

For recording we used a camera which is commonly employed in machine vision tasks. This camera was connected via IEEE 1394 interface (also known as FireWire) with the computer, so that all videos could be recorded digitally without the need of any frame grabber. The main reason for choosing a machine vision camera instead of a common television camera was that we were able to program our own recording software. Our software allows to control the camera settings and ensures an almost full automatic capturing of the sign language corpus. Further post-processing work was thus reduced to a minimum.

All videos were recorded directly onto hard disk using an image resolution of 776×578 pixels at 30 fps. This high spatial resolution ensures reliable extraction of manual and facial features from the same input image. For quick random access to individual frames, each video clip was stored as a sequence of images.

4.4. Recording Procedure

The reference signer's performance of the corpus was recorded first. His videos are thus called reference videos in the following. In order to ensure that all signers perform the same dialect, a reference video and its textual representation were prompted on a screen mounted below the camera. The reference video was shown once before recording started. After that the video vanished and only the text remained visible. When the camera started recording, the signer performed the prompted isolated sign or continuous sentence. If an error occurred, recording was interrupted by the supervisor and the performance was repeated.

4.5. Post-Processing

The video camera utilizes a single image sensor for the three primary colors red, green, and blue. For this reason the image sensor is covered by an array of color filters, also referred to as Bayer filter mosaic. Image sequences were captured in raw format first. Then each single image was post-processed as follows: Bayer demosaicing, vignetting removal, white balance correction, and image compression.

4.6. Resource Availability

The SIGNUM Database is available for academic and commercial use. In order to apply for a license, please contact one of the following distributors:

- Bavarian Archive for Speech Signals (BAS) ¹
- European Language Resources Association (ELRA) ²

For detailed documentation see (von Agris, 2009).

5. Experimental Results

The following experiments were carried out on the recorded SIGNUM Database. Recognition performance for isolated signs was evaluated using the 450 basic signs and for continuous signing using the 780 sentences. In both cases the evaluation of the signer-dependent (SD) performance is based on the three variations of the reference signer, whereas the signer-independent (SI) recognition rates were determined in a leave-one-out test on all 25 signers. Table 3 summarizes the experimental results.

		Vocabulary Size		
		150 signs	300 signs	450 signs
Isolated	SI	88.3%	84.5%	80.2%
	SD	96.0%	96.3%	96.9%
Continuous	SI	69.0%	68.4%	65.1%
	SD	87.5%	87.4%	87.3%

Table 3: Signer-independent (SI) recognition rates for isolated signs and continuous sign language. Rates for signer-dependent (SD) recognition are given for comparison.

The obtained results represent baselines without any adaptation. The classification stage was configured to employ neither subunit models nor any stochastic language model. As the corpus contains a high number of minimal pairs, the best recognition performance is obtained when both manual and facial features are exploited (von Agris et al., 2008b).

6. Conclusion

In this paper, we described the recording of the first sign language video corpus which meets the requirements for signer-independent continuous recognition. The corpus is based on a vocabulary of 450 basic signs in German Sign Language and comprises 780 sentences each performed by 25 native signers of different sexes and ages. The SIGNUM Database was made available for all interested researchers in order to establish the first benchmark.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (German Research Foundation). We thank Uwe Zelle for recording the sign language database.

7. References

- B. Bauer. 2003. *Erkennung kontinuierlicher Gebärdensprache mit Untereinheiten-Modellen*. Dissertation, Chair of Technical Computer Science, RWTH Aachen University.
- U. Canzler. 2005. *Nicht-intrusive Mimikanalyse*. Dissertation, Chair of Technical Computer Science, RWTH Aachen University.
- DESIRE. 2004. *Aachener Glossenumschrift. Übersicht über die Aachener Glossennotation*. Technical report, Deaf and Sign Language Research Team, RWTH Aachen University.
- G. Fang, W. Gao, X. Chen, C. Wang, and J. Ma. 2002. Signer-independent continuous sign language recognition based on srn/hmm. In *International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, pages 76–85. Springer.
- K.-F. Kraiss, editor. 2006. *Advanced Man-Machine Interaction*. Springer.
- S. C. W. Ong and S. Ranganath. 2005. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):873–891, June.
- U. von Agris, C. Blömer, and K.-F. Kraiss. 2008a. Rapid signer adaptation for continuous sign language recognition using a combined approach of eigenvoices, mllr, and map. In *Proc. of the 19th IAPR International Conference on Pattern Recognition*, Tampa, Florida, December.
- U. von Agris, M. Knorr, and K.-F. Kraiss. 2008b. The significance of facial features for automatic sign language recognition. In *Proc. of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, September.
- U. von Agris, J. Zieren, U. Canzler, B. Bauer, and K.-F. Kraiss. 2008c. Recent developments in visual sign language recognition. *Springer Journal on Universal Access in the Information Society*, 6(4):323–362, February.
- U. von Agris. 2009. *SIGNUM Database*. Sign language corpus, Online documentation. <http://www.phonetik.uni-muenchen.de/forschung/Bas/SIGNUM/>.
- M. Zahedi, P. Dreuw, D. Rybach, T. Deselaers, and H. Ney. 2006. Continuous sign language recognition - approaches from speech recognition and available data resources. In *Second Workshop on the Representation and Processing of Sign Languages*.
- J. Zieren and K.-F. Kraiss. 2005. Robust person-independent visual sign language recognition. In *Proc. of the 2nd Iberian Conference on Pattern Recognition and Image Analysis*, Lecture Notes in Computer Science.

¹<http://www.bas.uni-muenchen.de/forschung/Bas/BasSIGNUMeng.html>

²http://catalog.elra.info/product_info.php?products_id=1100

About Recognition of Sign Language Gestures

Alexander Voskresenskiy, Sergey Ilyin

School for deaf children (1), Academy of Fantasy (2)

Moscow, Russian Federation

E-mail: avosj@yandex.ru, mail@mocaprus.ru

Abstract

A motion capture technique for implementing sign language dictionary is described. Problems of perception and recognition of gestures of Russian sign language in system of the automated sign language translation are discussed. The new approach to morphology of gestures and a method for separate gestures in sign statements are offered. The working definition for "text understanding" is offered.

1. Introduction

The main goal of our work is to create Russia's first explanatory dictionary of Russian sign language, using three-dimensional animated characters created by motion capture techniques and training manuals that contain sign statements-assembled based on this dictionary.

The purpose of the vocabulary and tools is to help deaf people to learn the Russian verbal language, and promoting people who are learning sign language.

A fixed number of examples of verbal and sign statements contained in the manuals are not always able to meet the needs of the student. Therefore, in subsequent stages of work is supposed to create an automated system of sign language interpretation.

Currently there are no word processors, who understand text contents. Available word processors are based on statistical methods. This leads to a significant number of errors, reduction of which using the existing methods is hardly possible.

Our approach to the problem of understanding based on the fact that both verbal and sign language used to describe the same surrounding world. Therefore, we believe that the basic concepts describing the surrounding world for hearing and deaf people are the same.

A comparison of the meanings of words and gestures enabled us to formulate a working definition of the term "understanding of the text"¹.

Basic complexities at translation of text into signs are connected with homonymy resolution, searching of necessary meaning of polysemic word and/or sign, and also with transformation of phrases of Russian language into Sign Language expressions.

Procedure for transfer of sign statements in the text even more complicated, because the gesture utterance does not contain information about the grammatical forms of words from which to generate text, such as noun and verb in many cases are indicated by the same gesture.

The focus of this work is given to the separation of sign utterances into constituent gestures.

2. Current Results

2.1 Short Description of RuSLED Dictionary

Russian Sign Language Explanatory Dictionary RuSLED includes functions of explanatory dictionary as for entered word, so and for gesture representation. On input of dictionary any form of word can be entered, and at the output variants of gesture interpretation of given lexeme are shown.

Dictionary contains 2372 words (with interpretations of their meanings) and 2537 video images of gestures (including variants of the sign) which represent meanings of the words. For 1592 gestures (63% from total number in dictionary) additional explanatory, concerning to manner of execution of gesture or describing semantic nuances are given.

Gestures used in Saint Petersburg and its vicinities are presented in the dictionary. They in part coincide with Moscow gestures but divergence is big enough, what gave occasion to name given dictionary "Petersburg's dialect". In first version of dictionary digitized fragments of video recording borrowed from video course (IRRC, 2002) are used. Use for viewing of gestures of Windows Media Player ActiveX element allows: to see this gesture repeatedly, at pressing of button ► of player; to suspend performance of gesture in required place, at pressing of button || of player; to see any phase of gesture, moving cursor of player in appropriate position by mouse (fig. 1).

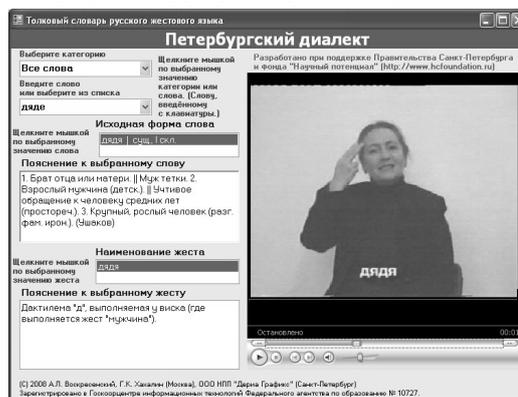


Figure 1: RuSLED dictionary display

¹ The term "text" is used to denote the means of exchange of knowledge between people, including both written text and speech communication (verbal and signed).

Video recording was used for best representation of mimicry accompanying gestures and executing essential role in sign language of deaf persons. So, for example, words «МИЛЫЙ», «СИМПАТИЧНЫЙ» (darling, nice) are passed with one gesture, but they are differing by movements of lips pronouncing fragments of corresponding words. In new dictionary version video records will be substituted by avatars using motion capture methods.

For some gestures explanatory from (Fradkina, 2001) were used. This dictionary is made on basis of Moscow variant of Russian Sign Language.

For compiling of words explanatory more 30 dictionaries and encyclopedias were used.

On deaf children teacher's recommendations opportunity is provided to filtration of word list of dictionary on grammatical categories (nouns, verbs, adjectives, adverb, pretexts, particles, numerals, pronouns). For viewing all dictionary content it is necessary to choose category "All words".

Separate input of dictionary (separate recording in table of database) is used for everyone semantic value of lexeme (and gesture). This dictionary feature is very convenient for user, and is recommended by lexicographers.

Field «Введите слово» ("Enter Word") allows to enter any word forms or choose lexemes from associated list. In list «Исходная форма слова» ("Initial Word Form") a lexeme corresponding to stem in field "Enter Word" is outputted or several lexemes are outputted if several records are chosen by results of morphological analysis.

When user chooses a lexeme from list "Initial Word Form" as result name of corresponding gesture is outputted in list «Наименование жеста» ("Name of Gesture"). If several gestures correspond to given lexeme then list of names of gestures is outputted. For each word meanings only that gesture is outputted, semantics of which corresponds to meaning of chosen lexeme (Voskresenskij & Khakhalin, 2007).

2.2 Our Approach to Understanding of Text

Word processing is usually divided into successive stages of morphological, syntactic, and, as a final stage, the semantic analysis. However, in some cases, morphological analysis can be performed only on the basis of syntax; in turn unambiguous parsing proposal assumes knowledge of grammatical forms of words in the sentence. Therefore consistent scheme of sentence parsing should be replaced by a scheme of interaction of agents performing different tasks and share the results to refine their work (Majumdar et al., 2008).

Modern systems for semantic text processing for removal of polysemy use ontology and thesauri. As the evaluation of the quality of such systems, the number of errors even in the best samples does not fall below 30% (Loukachevitch, Chuiko, 2007). The main reasons for this are incomplete vocabulary and inadequate procedures for resolving polysemy.

But what is the understanding of the text? The following definition was developed on the basis of comparison and

analysis of interpretations of the meanings of words and gestures:

The result of understanding of the text should be the selection and identification of objects described in the text, their spatial positions, as well as registration of changes to their characteristics, actions and conditions in accordance with the change of the text time.

According to the results for each given moment of the passed time of the text we can construct a picture, describing the locations and interactions of the objects are described in the text — the situation. In addition, the interpolation of changes of objects characteristics can provide short-term forecasting of changes of situations.

Supporting examples can be found in the RuSLED dictionary. Some of them are described in (Voskresenskij et al., 2009).

System of the text understanding should not only store information about semantic relationships of words (often ambiguous), defined by thesauri and ontology, but also must be able to speculate on the possible actions of the subject and the objects described in the text.

Identification of objects includes not only the allocation of group names that describe a particular object, but also recognition that, if the object met earlier in the text; if the objects are the same whether they have the same names (Kazi, Ravin, 2000). For this system, described in (Voskresenskij, 2008), includes not only the basic ontology, storing descriptions of classes and their relations, but also the ontology of the text, including descriptions of specific instances of classes. This ontology will inherit from the basic ontology characteristics of the classes and their relationships, adding to them the characteristics of specific instance (including its position in space).

For example, if the text describes the room in which there are several tables, then to understand what is at some of the tables, not enough to know a general description of the semantic class "table", each instance must be identified. But some of the hallmarks of an instance of a class can be meaningful only within a particular text, so they should not be included in the basic ontology. If they are repeated in different texts and for different instances, these features are important not only within a particular text. Then in the process of system self-learning they must be included in the basic ontology, leading to partition the source class into subclasses.

From this it follows that the ontology of a particular text should not be destroyed upon completion of the text processing, but should be kept for some time. It is necessary to compare information from different texts and identify the most plausible, which may be included in the basic ontology of the system.

The proposed approach to the understanding of the text is useful not only for sign language interpretation, but also for machine translation systems for verbal language. For example, in the Ingush language to convey information about the event, which ended recently, and in which the telling the story subject was present or absent, different forms of the verb are used.

3. Tools for mapping, perception and recognition of gestures

Various versions of the notation, for example Hamburg notation system HamNoSys², used to record gestures. In our country, notation proposed by L.S. Dimskis (2002) is used.

Dictionary of Russian sign language RuSLED is added by the function of gesture search using his approximate description. The challenge is that we need to find a gesture that people saw, but does not know its meaning. It uses a simplified notation, hided inside the dictionary, user-accessible lists of possible values: text to describe the place of performance gesture, text with a pattern — for the configurations of fingers. Based on user-selected values search query is formed and returns a set of gestures to meet this request, from which the user selects the gesture.

Demonstration of gestures in the new version of the dictionary made by animated character - an avatar, to record of gestures method of motion capture is used. Record is performed by "The Academy of Fantasy» (www.mocaprus.ru). Movements of demonstrator recorded using 12 cameras and a host of reflectors on the suit (Fig. 2), are converted to 3D-model (Fig. 3), and used to form the shape of an avatar that can be placed into any stage.

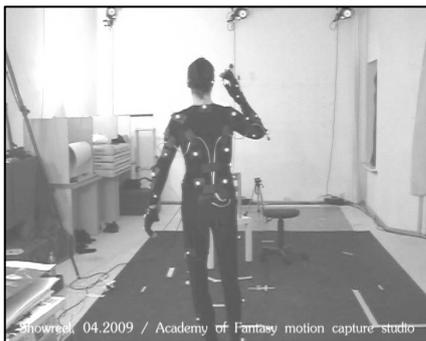


Figure 2: Demonstrator in a suit with reflectors



Figure 3: 3D-model

Movements of the fingers of the demonstrator are recorded using special gloves. To record the facial expressions and articulation the apparent on the face of the reflectors is used (Fig. 4). Their signals are converted into three-dimensional model of facial mimicry (Fig. 5).



Figure 4: Demonstrator with glued on the face reflectors

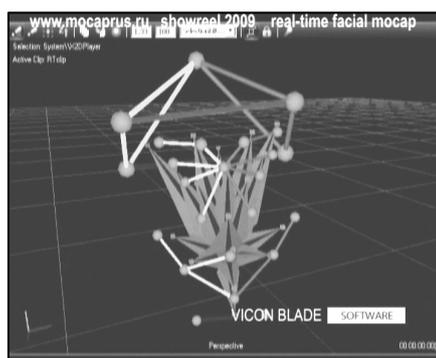


Figure 5: Three-dimensional model of facial mimicry

This will significantly expedite the filling of the dictionary through the use of several sign language interpreters for the demonstration of gestures, while preserving the unity of action expressed by the appearance of a single virtual gesture demonstrator. Formed in such a way dictionary will allow composition of sign statements of gesture collections stored in the dictionary maintaining, as noted above, the unity of action which is important for perception of sign utterances by human.

Studio recording of gestures, allowing you to create original dictionary, obviously, can not be a means of communication with deaf people.

For recognition of gestures there is proposed to develop means of converting raster images of sign language interpreter taken with a camera in a vector images. This transformation includes the recognition of the essential for this task image detail: the head, hands (and the position of each finger), the torso. These details of the image are converted to ellipses and rectangles, the coordinates of which are compared with the parameters of the skeleton of a virtual demonstrator (avatar) of the dictionary.

Transformation perceived image in vector form allows you to significantly reduce the memory requirements of the intellectual system and accelerate the procedure of comparison with etalons.

Methods for converting the image to be used are similar to those used in the pre-processing of images in the systems of character recognition.

Information on the exact position of avatar in space, such as hands, which was absent in two-dimensional scanning

² <http://www.sign-lang.uni-hamburg.de/projects/hamnosys.html>.

images from a camcorder, it is planned to receive from the knowledge of possible and permissible mutual positions of various parts of the body. To determine the exact pose of avatar the appropriate geometric constructions will be applied, providing the closest match with the original raster image the projections of the avatar on the plane.

4. Methods of processing of sign phrases

In the analysis of sign utterances should take into account that many of the signs are composite, contain a combination of several gestures and pre dactyl signs, modifying the meaning of this sign. When you need to specify, for example, case endings, after a gesture relevant dactyl signs are signed.

Gestural speech does not contain pauses between individual gestures. Only phrases are separated by pauses. This introduces additional complexity in the implementation of automated sign language translation, like those encountered in the development of continuous speech recognition systems.

Given the integral nature of the gestures, the separation of gestural phrases into separate gestures should be maintained by selecting from the vocabulary appropriate gestures, having the greatest length, and analyzing the semantics of the resulting expressions. If its meaning does not match the discourse, we can proceed to successively splitting "long" sign on the constituent elements, trying to get a statement, the content of which corresponds to the discourse. Considering also that the gesture might pass the words of different grammatical forms, construction of syntax tree of a text sentence offers a complex combinatorial problem whose solution is a simple brute force attack is impossible, since it leads to the "exponential explosion".

The solution is to use the method of sequential analysis and retention options without the incremental construction of solutions (Mikhalevich, Volokovich, 1982), which reduces the number of options under consideration. This criterion for excluding unpromising options is contradictory semantics of the resulting text.

5. Conclusion

In the case of sufficiently reliable recognition of gestures using a camcorder (preferably a qualitative recognition using standard web cameras) and establishing a system of sign language interpretation will be possible to ensure prompt communication of the deaf with administration officials and the public, that is a function of "electronic government".

Many details of the process of understanding and explanation, expressed in words, hidden from direct observation in the subconscious, which hampers the development of word processors understanding the text. Based on a comparison of different models of thinking, presented verbally and in sign language, developed a model for understanding the text. Accordingly, there is an idea of the architecture of a system that could perform the required functions.

6. Acknowledgements

The authors express sincere gratitude to N.A. Chaushian. Her very useful comments on sign language allowed drawing attention to the features not always visible to researcher who is not a native signer, and correcting some errors in dictionary RuSLED.

7. References

- Dimskis, L.S. (2002). *Start Learning Sign Language*. Moscow: Academy. (In Russian: Димскис Л.С. Изучаем жестовый язык. М.: Издательский центр «Академия», 2002. — 128 с).
- Fradkina, R.N. (2001). *Hands Which Are Talking*. Moscow: All-Russian Deaf Association. (In Russian: Фрадкина Р.Н. Говорящие руки: Тематический словарь жестового языка глухих России. М.: Изд-во ВОИ. — 598 С).
- IRRC, (2002). Specific communication means of deaf : videocourse in 3 parts. Saint-Petersburg – Pavlovsk : Inter-regional Rehabilitation Center. (In Russian : Специфические средства общения глухих: Видеокурс: В 3 частях. СПб – Павловск: МЦР, 2002).
- Kazi, Z., and Y. Ravin. (2000) Who's Who? Identifying Concepts and Entities across Multiple Documents. *Proceedings of the 33rd Hawaii International Conference on System Sciences*. (0-7695-0493-0/00).
- Loukachevitch, N., and D. Chuiko (2007). Thesaurus-based Word Sense Disambiguation (In Russian: Лукашевич Н.В., Чуйко Д.С. Автоматическое разрешение лексической многозначности на базе тезаурусных знаний. *Интернет-математика-2007*. Екатеринбург: Изд-во Урал. ун-та. — С. 108 – 117).
- Majumdar, A., J. Sowa, and J. Stewart. (2008) Pursuing the Goal of Language Understanding. *Proceedings of the 16th ICCS / P. Eklund and O. Hammerlé, eds. LNAI 5113, Berlin: Springer, pp. 21-42.*
- Mikhalevich V.S., Volkovich V.L. (1982) *Computational methods of research and design of complex systems*. Moscow: Nauka. (In Russian: Михалевич В.С., Волкович В.Л. Вычислительные методы исследования и проектирования сложных систем. М.: Наука, 1982. — 288 с).
- Voskresenskij, A. (2008). Text Disambiguation by Educable AI System. *The First Conference on Artificial General Intelligence / P. Wang et al. (Eds.). Amsterdam: IOS Press. — P. 350 – 361.*
- Voskresenskij, A., and Khakhalin, G. (2007). Semantic Search Engine in a Multimedia Russian Sign Language Dictionary. *Proceedings of the XIIIth International Conference "Speech and Computer" SPECOM'2007, Moscow, Russia, pp. 739 – 744.*
- Voskresenskiy A.L., Gulenko I.E., and Khakhalin G.K. (2009) From Sounding Speech to Sign Language. *Proceedings of the 13-th International Conference "Speech and Computer" SPECOM'2009. St. Petersburg: SUAI, 2009. P. 539 – 542.*

Sign Languages of Europe – Legal Status and Human Rights

Mark Wheatley, Annika Pabsch

European Union of the Deaf
Rue de la Loi/Wetstraat 26/15, 1040 Brussels, Belgium
E-mail: mark.wheatley@eud.eu, annika.pabsch@eud.eu

Abstract

Sign languages across the globe are fully-fledged languages that differ between Deaf communities throughout Europe and the world. A recent survey by the European Union of the Deaf gathered that there are about 650,000 sign language users in the EU for whom using a sign language is the only way to communicate and have equal access. It is therefore crucial to legally recognise national sign languages. Being treated equally without prejudice also with regards to language is a basic Human Right as postulated in the UN Declaration of Human Rights. Other rights, such as the right to education and a fair trial can only be guaranteed if sign languages are recognised as distinct languages in order to provide sign language interpreters and education in sign language. At EU level, a number of documents and Resolutions have been adopted but so far only three European countries have recognised sign language at constitutional level: Austria, Finland and Portugal. Other countries, such as Hungary and Spain have taken other legal measures to protect their sign languages. Although Europe's sign languages enjoy some recognition, sign language users across Europe are still lacking legal protection at the same level as other minorities.

1 European Union of the Deaf (EUD)

The European Union of the Deaf (EUD) is a European non-profit making organisation whose membership comprises National Associations of the Deaf (NAD). Established in 1985, the EUD is the only organisation representing the interests of the Deaf¹ at European Union level. The mission's aim is to promote and advance the (Human) Rights of the Deaf in Europe by achieving the recognition of the right to use sign languages, empowerment, and equality in education and employment (EUD 2010a).

2 Sign Languages

2.1 National Sign Languages

Despite widespread opinions there is not one single universal sign language in the world. Sign languages vary between countries and ethnic groups; some countries even have two or more sign languages such as Belgium or Switzerland. Sign languages also show distinct dialects that vary from region to region. Nonetheless, national sign languages are fully-fledged languages that have a grammar and lexicon just as any spoken language (see for example Sutton-Spence & Woll, 1999 for British Sign Language).

2.2 International Sign

At international conferences or meetings an auxiliary language – referred to as International Sign (IS) – is used to communicate among Deaf people who do not share a common language. The EUD uses IS for example at board meetings and its Annual General Assembly, as determined in the EUD Internal Rules (EUD 2010c). The World Federation of the Deaf (WFD) has even established it as its official working language next to English (WFD 2003). IS is however not actually a discrete language, it is a contact language whereby signers will use signs from their respective natural sign languages along with established IS signs and simplified grammatical structures (Locker McKee & Napier 2002).

2.3 EUD survey

A recent EUD survey (2008) estimated that there are about 650,000 sign language users in the EU. This is not to be confused with the number of deaf or hard of hearing people, which is much higher. Currently, there are approximately 7,000 sign language interpreters in the EU. This results in an average ratio of 93 sign language users to 1 sign language interpreter. Among the EUD members, Finland has the highest ratio of 6 to 1 and Slovakia the lowest with 3,000 to 1.

Although Finland's ratio of 6 to 1 might sound very good compared to other European countries, this is still not enough to provide for all Deaf people. There is no "ideal" ratio that could be stated here but it is worth noting that although Finland's number is close to ideal, the profession of Finnish Sign Language interpreters is not adequately paid. A Finnish Sign Language interpreter earns for example €18.45/hour (SVT 2010). In comparison, a British Sign Language Interpreter charges about €23/hour on average (ASLI 2008). In most countries, sign language

¹ Deaf with a capital 'D' relates to Deaf people who consider themselves part of the Deaf Community and use sign language as their first or preferred means of communication. This is in contrast to 'deaf', which merely describes the audiological status of non-hearing.

interpreters are paid less than spoken language interpreters and have a lower professional status. The US seems to be one of the only countries where this is not the case (Bancroft, 2005). So although Finnish Deaf people might have more access than Slovakian Deaf people, the standard is by no means the same as that of a hearing person or even a member of a minority group receiving interpretation into their mother tongue.

3 Sign Language and Human Rights

3.1 Sign Language as Mother Tongue

Sign languages are the only languages Deaf people are able to acquire naturally and spontaneously (Jokinen, 2000). Therefore, they should be seen as the mother tongue of Deaf people, although most Deaf people (approximately 90%) grow up in hearing families and do not necessarily learn a sign language until later in life (Krausneker, 2006). Sign language as the mother tongue of Deaf people is in accordance with EUD's philosophy and also with Skutnabb-Kangas and Bucak's (1995) external definition of mother tongue, which states that a person must identify with his/her language and/or use it as a primary means of communication. This means that although the respective sign language might not be acquired in the family and also not necessarily as the first language (L1), it should nonetheless be treated as the mother tongue of Deaf people – not only in educational settings but also regarding access to work and public authorities.

When a Deaf child has been granted the right to its mother tongue it is also later more likely to be able to learn the surrounding majority language (in its written form) and this will furthermore increase chances to have access to higher education or other further education programmes (Emery 2009).

3.2 Human Rights

The UN Declaration of Human Rights (UNDHR 1948) grants rights to everyone regardless of certain characteristics such as language or religion. A person is to be treated equally, even if he or she does not speak the national language. For minority language speakers this becomes an issue if their language is not protected by legal measures. It also means that although everyone should be treated equally, this person will receive additional services – such as a (sign) language interpreter – if the language is legally recognised. This is for example the case with Welsh in the UK. It is argued that sign languages should be legally recognised to grant Deaf people Human Rights with regards to their language, as the language is the key to other basic Human Rights, such as education or fair trial.

3.3 Linguistic Human Rights

Using Skutnabb-Kangas and Bucak's (1995) definition of mother tongue, Deaf people can claim Linguistic Human Rights (LHR) in regard to sign language (Skutnabb-Kangas 2000). LHR are a hypothetical concept but in

recent years legislation in the EU and the world has come into effect giving Deaf people more and more rights with regards to sign language and equal access. Skutnabb-Kangas claims that LHR are language rights that are needed to guarantee basic Human Rights. For example, in order to gain access to education, a person needs to be able to understand the teacher. This is only possible when having primary education in one's mother tongue. Additionally, she states that education should not only be in the medium of the mother tongue but that the language should also be taught as a subject in schools. She also grants collective rights to minority groups, such as the right to exist. Using this theoretical concept is useful in understanding the (Human) Rights that Deaf people are denied on a daily basis when not being able to use their language with authorities, in trials, at school or at work.

3.4 Minority Rights

Although d/Deaf people are often only seen as a disability group in need of support, Deaf people see themselves as a minority group with a distinct language. Just as any other member of a minority group, Deaf people require access in their mother tongue. It is even more crucial for Deaf people to be granted this right, as they are not physically able to learn spoken languages to a level that is sufficient to communicate with hearing people directly. Currently most legislation relating to sign language and sign language interpreter provision is embedded in disability legislation. Although this is sometimes seen as not fully recognising a national sign language, it is an effective means to provide for access.

4 Sign Language Legislation

4.1 UN Convention

The recent UN Convention on the Rights of Persons With Disabilities (UNCRPD), which came into force in 2008, is the first international document to mention sign language explicitly. It is a milestone in achieving Human Rights for Deaf people, as it grants rights concerning accessibility and education. It places legal obligations on States to abolish discrimination and protect and promote the rights of persons with disabilities, including Deaf people. The UNCRPD requires States to take measures to provide assistance for example in the form of professional sign language interpreters.

Although this is a first step it is questionable what effect the Convention will have in the near future as individual States as well as the EU have adopted the Convention but not yet widely implemented it in respective country legislation.

4.2 EU Resolutions

Apart from adopting the UNCRPD, the European Parliament adopted a Resolution on Sign Languages in 1988, which was reiterated in 1998. It calls on the European Commission and its member States to legally recognise the sign languages of Europe. The Resolution also acknowledges the fact that a number of Deaf people

need to communicate in sign language, as this is the only possible language. Moreover, it states that the sign languages of Europe are distinct languages that each have their own cultural identity.

These two Resolutions are not legally binding but show that the European Parliament is aware of the needs of sign language users across Europe. The fact that this Resolution was already adopted over 20 years ago but the situation of Deaf people has – in some countries – not changed significantly is worrying and makes it clear that a Resolution at EU level might not be the best way to give Deaf people equal Human Rights. To achieve that European instruments adequately protect Deaf people, the EUD works closely with representatives of the European Parliament.

4.3 Council of Europe

The Council of Europe (CoE) has published a number of reports and recommendations regarding sign languages in its member states. Most notably it published a Recommendation regarding the protection of sign languages in member States in 2003 (Rec 1598). This recommendation takes note of an older Recommendation (Rec 1492) relating to minority languages including sign languages. Although not legally binding, such a document shows the CoE's growing awareness of a need to protect sign languages in the same way as other minority languages.

Krausneker (2008) submitted an expert opinion for the CoE regarding the needs of sign language users across Europe. This needs analysis does not only offer concrete examples of how to tackle inequalities, the paper also describes clearly how sign language users need access in sign language to be granted full (Linguistic) Human Rights. It formulates 25 recommendations that States should adopt and implement. These range from legal recognition of sign languages as part of minority rights to granting access to information.

5 Sign Language as a Constitutional Right

5.1 Sign Language Legislation

Although sign languages have been recognised at EU and UN level to a certain extent it is important for individual countries to change their laws accordingly. If this is not done there is a risk of these laws not having any effect on Deaf people's life. Only three countries have recognised their national sign language at constitutional level: Austria, Finland and Portugal².

Some countries – such as Finland, Spain the Czech Republic or Slovakia – have passed laws that give Deaf people rights with regards to education, sign language interpreters, or access to work. Hungary recently (2009) adopted the most comprehensive piece of sign language

legislation granting – among other things – the right to learn Hungarian Sign Language and have access in that language through a free interpreter service that is funded by the State. This is a significant and unique piece of legislation, as it immediately provided the funds necessary for the free interpretation service in its current Budget Act and added specific deadlines for example for the provision of in-vision interpreters and subtitling on national television. This shows that although Hungary has not recognised its sign language at constitutional level rights are nonetheless accorded to Deaf people. It also makes clear it is not necessarily better for the provision of services to recognise the national sign language on such a high level. Sweden has for example incorporated a bilingual education policy in their education law, which has had a greater effect in Deaf people's lives than a single sentence recognising ÖGS (Austrian Sign Language) in the Austrian Constitution, as Verena Krausneker noted at the EUD seminar in 2009.

5.2 Austria

Austria recently (2005) changed its Federal Constitution to contain an Article on ÖGS. It states: "Austrian Sign Language is recognised as a fully-fledged language. More shall be regulated by further laws" (Article 8(3)). This is a significant step for Deaf people in Austria, although no further laws have thus far been enacted. The positioning of the sentence is also an important factor as paragraph (1) and (2) deal with the national language of Austria and the linguistic and cultural diversity of the country.

The law has had an effect on the teacher training and on educational policy in general. Austrian Sign Language is now part of teacher training and although no formal law has been passed, a certain shift in attitude can be seen. For example a number of Deaf schools in Vienna now have adopted bilingual education policies.

5.3 Finland

Finland was the first European country to recognise sign language at constitutional level in 1995. The Constitution of Finland states: "The rights of persons using sign language and of persons in need of interpretation or translation aid owing to disability shall be guaranteed by an Act" (Chapter 2 Basic rights and liberties, Section 17 Right to one's language and culture). Finland has a history of being a country with two official languages – Finnish and Swedish – and granting equal rights to speakers of these two languages. Additionally, the Sami and Roma are given the status of an indigenous people. Sign language is mentioned in the same section as Sami and Roma, which gives it a similar status as these minority languages.

It is also worth noting that the section does not speak of deaf people but of "persons *using* sign language". This is significant, as not everyone who is deaf necessarily is in need of a sign language. But on the other hand the Constitution does not recognise Finnish or Swedish Sign Language as a specific language, like it was done in Austria. Contrary to Austria Finland has however adopted a range of other pieces of legislation, which further regulate the recognition of sign language. Most notably,

² It should be noted that some countries, such as the UK do not have a Constitution and therefore cannot legally recognise sign languages at such a level.

the Finnish Education Act (628/1998) recognises sign language as a mother tongue that needs to be taught as well as being used as the language of instruction. Finnish Sign Language is interestingly also mentioned as a requirement for naturalisation if oral skills cannot be demonstrated (see Nationality Act 359/2003). Other Acts include the Language Act and the Act on Yleisradio Oy, the Finnish Broadcasting Company. The Finnish Parliament is also currently (Spring 2010) discussing a new legislative proposal for interpreting services for persons with disabilities, which aims to ensure at least 180 hours of free interpreting services per year, excluding educational needs (EUD 2010b).

The example of Finland shows that when sign language is recognised at constitutional level, Acts need to follow to have an effect on Deaf people.

5.4 Portugal

Portugal recognised Portuguese Sign Language in 1997 in its Constitution. Article 74(1)2 on education states: “In implementing the education policy, the State shall be responsible for [...] h) Protecting and developing Portuguese Sign Language as an expression of culture and an instrument for access to education and equal opportunities”. Although mentioning Portuguese Sign Language and not only the term sign language, as seen in the Finnish Constitution, it is significant that the language is recognised in the article relating to education, which shows that Portugal does not see their national sign language as the mother tongue of Deaf people, or as a minority language; it is seen as an “instrument”. But on the other hand it means that education is provided in sign language. Portugal has not – like Austria – adopted other legal measures to provide for example access to work through a free sign language interpretation service.

6 Summary

The recent EUD survey (2008) investigated sign language use and legislation in Europe to gain a clearer picture of the current legal situation. Three countries have recognised sign language at constitutional level. Although recognising that this is an important step for sign language users in the respective countries, legal recognition at such a high level has to be seen with caution. Formally, it is an improvement but in reality it sometimes has no or only little effect. Deaf people are however in need of a legal basis to defend their basic Human Rights. The recent UNCRPD grants these rights but nonetheless, this important document needs to be implemented in the relevant country legislation to have an effect in Deaf people’s lives. Interpreters are not yet widely available and education is often still provided orally rather than using a bilingual approach. When having put legislation in place, it is crucial to then provide financial means and ensure these services can actually be provided by for instance fostering interpreter training programmes. Overall, there are various pieces of legislation in place but it cannot be forgotten that sign languages are minority languages that their speakers depend on, as they have no equally efficient means of communication. This makes

recognition crucial and lets this issue become a true question of Human Rights.

7 References

- Association of Sign Language Interpreters ASLI (2008). Fees and Salaries Report. Available at: <http://www.asli.org.uk/fees-salaries-report-p122.aspx> (Accessed on 23 March 2010).
- Bancroft, M. (2005). *The Interpreter’s World Tour – An Environmental Scan of Standards of Practice for Interpreters*. Woodland Hills, CA: The California Endowment.
- Emery, S.D. (2009). In Space No One Can See You Waving Your Hands: Making Citizenship Meaningful to Deaf Worlds. In *Citizenship Studies*, 13(1), pp. 31-44.
- European Union of the Deaf EUD (2010a). EUD. Available at: <http://eud.eu/EUD-i-14.html> (Accessed on 23 February 2010).
- European Union of the Deaf EUD (2010b). Finland. Available at: <http://eud.eu/Finland-i-182.html> (Accessed on 20 March 2010).
- European Union of the Deaf EUD (2010c). Internal Rules. Unpublished Document.
- European Union of the Deaf EUD (2008). Unpublished survey.
- Jokinen, M. (2000). The Linguistic Human Rights of Sign language Users. In R. Phillipson (Ed.), *Rights to Language: Equity, Power and Education*. New York: Lawrence Erlbaum Associates, pp. 203-213.
- Krausneker, V. (2008). Report on the protection and promotion of sign languages and the rights of their users: needs analysis. Council of Europe Publishing.
- Krausneker, V. (2006). *Taubstumm bis Gebärdensprachig – Die österreichische Gebärdensprachgemeinschaft aus soziolinguistischer Perspektive*. Meran: Alpha & Beta Verlag.
- Locker McKee, R. & Napier, J. (2002). Interpreting into International Sign Pidgin – An Analysis. In *Sign Language & Linguistics* 5(1), pp. 27-54.
- Skutnabb-Kangas, T. (2000). Linguistic genocide in education - or worldwide diversity and human rights? Mahwah, NJ: Erlbaum Associates.
- Skutnabb-Kangas, T. and Bucak, S. (1995). Killing a Mother Tongue – How the Kurds are Deprived of Linguistic Human Rights. In T. Skutnabb-Kangas, R. Phillipson (Eds.) *Linguistic Human Rights: Overcoming Linguistic Discrimination*. New York: Mouton de Gruyter, pp. 347-370.
- Sutton-Spence, R. & Woll, B. (1999). *The Linguistics of British Sign Language – An Introduction*. Cambridge: University Press.
- Suomen Viittomakielen Tulkit SVT (2010). *Interpreting Services in Finland* Available at: <http://www.tulkit.net/in-english/interpreting-services-in-finland-2006/> (Accessed on 23 March 2010).
- Wessel, M. (2003). *Sign Language of the Deaf as Minority Language*. Unpublished EUD paper.
- World Federation of the Deaf WFD (2003). *Statutes*. Available at: www.wfdeaf.org/pdf/statutes.pdf (Accessed on 18 March 2010).

Author Index

Almohimeed, Abdulaziz	7	Hong, Sung-Eun	178
Athitsos, Vassilis	11	Hoyoux, Thomas	65, 192
Auer, Eric	150	Hrúz, Marek	41
Badia, Toni	154	Huenerfauth, Matt	121
Balvet, Antonio	15	Ilyin, Sergey	247
Bardeli, Rolf	150	Ivanova, Nedelina	125
Bayley, Robert	98	Jantunen, Tommi	129
Bertoldi, Nicola	19	Jennings, Vince	133
Bordag, Stefan	150	Johnston, Trevor	137
Borgotallo, Roberto	23	Kalimeris, Constandinos	76
Bowden, Richard	57, 80	Kanis, Jakub	41
Braffort, Annelies	80, 88, 158	Karioris, Panagiotis	158
Branchini, Chiara	98	Kennaway, Richard	84, 133
Brashear, Helene	27	Konrad, Reiner	178
Buehler, Patrick	33	König, Lutz	106
Bueno, Javier	84	König, Susanne	178
Camp, Pavel	41	Kraiss, Karl-Friedrich	243
Cardinaletti, Anna	53, 98	Krňoul, Zdeněk	143
Cavender, Anna	45	Ladner, Richard	45
Cecchetto, Carlo	98	Langer, Gabriele	178
Cherniavsky, Neva	45	Langer, Jiří	41
Choisier, Annick	158	Lee, Seungyon	27
Chon, Jaehong	45	Leeson, Lorraine	116
Collet, Christophe	49, 80	Lesmo, Leonardo	19
Conte, Genny	53	Lillo-Martin, Diane	112
Cooper, Helen	57	Lombardo, Vincenzo	19
Crasborn, Onno	61, 65, 92, 186, 212	López, Verónica	208
Damiano, Rossana	19	Lu, Pengfei	121
Damper, Robert	7	Machado Oliveira, Carlos R.	147
Dandapat, Sandipan	172	Maragos, Petros	80, 196
Del Principe, Andrea	19	Marino, Carmen	23
Delorme, Maxime	88	Martinez, Gregorio	65
Dimiou, Nassia	158	Martín, Raquel	208
Dimou, Athanasia-Lida	76	Mascret, Bruno	168
Donati, Caterina	98	Masneri, Stefano	150
Dreuw, Philippe	65, 225	Massó, Guillem	154
Du, Wei	65, 192	Mathur, Gaurav	112
Duarte, Kyle	73	Matsusaka, Yosuke	231
Efthimiou, Eleni	76, 80, 102, 158	Matthes, Silke	106, 158
Elliott, Ralph	84, 133	Mazzei, Alessandro	19
Everingham, Mark	33	McDonald, John C.	217
Filhol, Michael	88	Mereghetti, Emiliano	98
Forster, Jens	65, 92, 186, 225	Metaxas, Dimitris	164
Fotinea, Stavroula-Evita	76, 80, 102	Michael, Nicholas	164
García, Adolfo	208	Milachon, Fabien	49
Geraci, Carlo	53, 98	Moreau, Cedric	168
Gibbet, Sylvie	73	Morrissey, Sara	172
Gilchrist, Shane	172	Moya Lazaro, José Miguel	65, 221
Giudice, Serena	98	Müller, Luděk	41
Glauert, John	80, 84, 133, 204	Nakazono, Kaoru	231
Gonzalez, Matilde	49	Nash, Joan	11
Goslin, Kyle	116	Nauta, Ellen Yassine	186
Goudenove, François	80	Neidle, Carol	11, 164
Goulas, Theodore	102	Ney, Hermann	65, 92, 225
Gumiel, Jesús	221	Nishio, Rie	178
Gweth, Yannick	65	Nolan, Brian	116
Hamilton, Harley	27	Nunnari, Fabrizio	19
Hanke, Thomas	80, 106, 110, 158, 178	Ormel, Ellen	65, 92, 186
Hochgesang, Julie A.	112	Pabsch, Annika	251
Hofmann, Markus	116	Pelhate, Julia	158

Piater, Justus	65, 192	Tanaka, Saori	231
Piccolo, Elio	19, 23	Thangali, Ashwin	11
Pissaris, Michalis	102	Theodorakis, Stavros	196
Pitsikalis, Vassilis	196	Thorvaldsdottir, Gudny Bjork	235
Poletti, Fabio	98	Tiotto, Gabriele	19, 23, 239
Presti, Peter	27	Tschöpel, Sebastian	150
Prinetto, Paolo	19, 23	van der Kooij, Els	186
Rathmann, Christian	178	van Dijken, Lianne	186
Riskin, Eve	45	Vanam, Rahul	45
Rossini, Mauro	23	Vendrame, Mara	239
Safar, Eva	158, 204	Verges Llahi, Jaume	65
San-Segundo, Rubén	208	Villanueva, Pedro Pascual	112
Santoro, Mirko	53, 98	von Agris, Ulrich	243
Schembri, Adam	212	Voskresenskiy, Alexander	247
Schneider, Daniel	150	Wagner, Sven	106, 110
Schnepp, Jerry	217	Wald, Mike	7
Schreer, Oliver	150	Wang, Haijing	11
Sciaroff, Stan	11	Wheatley, Mark	65, 251
Serrano, Marina	221	Wittenburg, Peter	150
Sheikh, Haaris	116	Wobbrock, Jacob	45
Sloetjes, Han	61, 150	Wolfe, Rosalee	217
Smith, Robert	172	Yuan, Quan	11
Somers, Harold	172	Zafrulla, Zahoor	27
Sánchez, David	208	Zelle, Uwe	225
Starner, Thad	27	Zisserman, Andrew	33
Stefan, Alexandra	11	Zucchi, Sandro	98
Stein, Daniel	65, 92, 186, 225	Železný, Miloš	41
Storz, Jakob	110, 158		