A German Sign Language Corpus of the Domain Weather Report

Jan Bungeroth, Daniel Stein, Philippe Dreuw, Morteza Zahedi, Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University
52056 Aachen, Germany
{bungeroth,stein,dreuw,zahedi,ney}@cs.rwth-aachen.de

Abstract

All systems for automatic sign language translation and recognition, in particular statistical systems, rely on adequately sized corpora. For this purpose, we created the Phoenix corpus that is based on German television weather reports translated into German Sign Language. It comes with a rich annotation of the video data, a bilingual text-based sentence corpus and a monolingual German corpus.

1. Introduction

The automatic recognition and statistical machine translation of sign language are challenges for the field of natural language processing. Systems capable of such processing are especially useful to help deaf people to communicate with their hearing environment, as human interpreters are expensive and not always available. Furthermore, (Traxler, 2000) shows that the majority of the deaf society have only poor to moderate reading skills. This makes automatic aids even more valuable.

A statistical system capable of such sign language processing (as described in Section 3.), needs a suitable corpus of sign language data to train with. Unfortunately, most of the currently available corpora are too small or too general for the mentioned tasks.

We therefore present a new corpus called Phoenix for the languages German and German Sign Language (DGS) for the restricted domain of weather reports. It comes with a rich annotation of video data, a bilingual text-based sentence corpus and a monolingual German corpus.

2. Related Work

Several groups worked on sign language corpora, but most of them focused on linguistic aspects rather than natural language processing:

- The European Cultural Heritage Online organization (ECHO)[1] published corpora for Swedish sign language, British sign language and the sign language of the Netherlands. All these corpora contain tales and stories each signed by a single signer. For our purposes they are too small and have a large vocabulary which makes automatic learning difficult. Though, in related work, (Morrissey and Way, 2005) applied example based methods for automatic translation based on one of the ECHO corpora. However, their results imply that their system is only robust for sentences already seen in training, but has problems with unseen word and phrase combinations.

- The American Sign Language Linguistic Research group at Boston University created a set of videos in American sign language which is partly available on their website[2] and described in (Neidle et al., 2000). All videos are annotated and recorded from three different perspectives. (Zahedi et al., 2005) published results on sign language recognition for this corpus. The corpus has focus on linguistic topics, though.

- (Heßmann, 2001) published a corpus based on interviews in DGS with several thousand sentences. This is publicly available on the ECHO website too. While this corpus is quite large, its domain is too broad, making automatic learning difficult.

3. System Overview

The complete sign language system as proposed by (Bauer et al., 1999) and (Bungeroth and Ney, 2004) is designed to translate written text into sign language and vice versa.

For the translation from German into DGS, a statistical machine translation (SMT) system is trained on a corpus consisting on glosses and German. Then, the German sentences can be automatically translated into the semantic representation (glosses) of a corresponding sign language sentence, and then converted into a syntatic notation (HamNoSys). This can be signed by an virtual character, the avatar.

For the direction from DGS into German, first the sign language should be recognized in their gloss notation. These glosses can then be translated into German again by the SMT system.

This brief summary of the system should emphasize the importance of appropriate corpora for both translation and recognition. Larger corpora of good quality improve the results when used for training.

4. Gloss Notation

For storing and processing sign language, a textual representation of the signs is needed. While there are several notation systems covering different linguistic aspects, we focus on the so called gloss notation. Glosses are widely used for transcribing sign language video sequences; they are a form of semantic representation for sign language.

In our work, a gloss is a word describing the content of a sign written with capital letters. Additional markings are

[1] http://www.let.kun.nl/sign-lang/echo/

[2] http://www.bu.edu/asllrp/

used for representing the facial expressions and other non-manual markings. The manual annotation of sign language videos is a difficult task, so notation variations within one corpus are often a common problem. To avoid this, we follow the specifications of the Aachener Glossenumschrift (DESIRE, 2004) in this work.
As an example, the following sentence is taken from the Phoenix corpus.

```
HOCH++ ATLANTIK WACHSEN-(mehr)-hn
```

It can be translated into English with ‘The high pressure areas over the Atlantic ocean are growing larger’. The three signs are transcribed with glosses ‘HOCH’, ‘ATLANTIK’ and ‘WACHSEN’ representing their meaning in German. Signs repeated (for example to indicate plural forms) are annotated with a double-plus, mouth pictures are written in brackets, e.g. ‘(mehr)’, ‘-hn’ means that the signer is nodding during signing.

5. Corpus Setup

The German television channel Phoenix broadcasts the daily news show Tagesschau in German and DGS. The DGS translation is provided by an interpreter who is shown in the lower right corner of the TV frame. The interpreter is changing daily. In total there are eight interpreters, some signing in different dialects of DGS, responsible for the translation of the news. Figure 1 gives an example image.

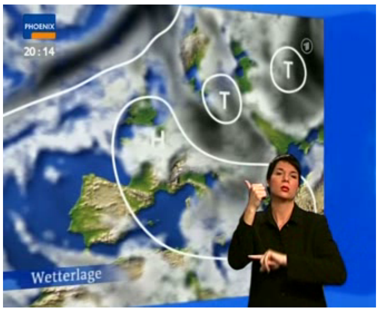


Figure 1: Example of the Phoenix corpus video data

For restricting the domain of the corpus, it was decided to concentrate on the weather reports only. This has several advantages:

- Similar sentences occur more often, as the structure of the weather reports does not change.
- The size of the DGS vocabulary is limited. However, the German vocabulary is less restricted and has about two times the size of the DGS part.
- Sign language specific grammatical structures like the use of space or facial expressions that are difficult to model, are less common in this domain.

The corpus which we name Phoenix consists of three parts: the annotated video files, the bilingual sentence corpus and a monolingual corpus for German. We introduce these parts in detail:

5.1. The Video Corpus

The news transmissions are recorded and converted to the MPEG1 video format which is necessary for the annotation software ELAN[3]. The weather report parts of the news are then annotated by a deaf DGS native speaker, and the quality of the annotations are checked regularly. In total 104 files are annotated. Figure 2 shows the ELAN software with a video file in the upper left corner and five annotation tiers at the bottom which are discussed below.
All Phoenix annotations are stored as EAF files, i.e. an XML format used by the ELAN software. This allows several annotations on different tiers on the same time line. Our annotation includes up to six different tiers:

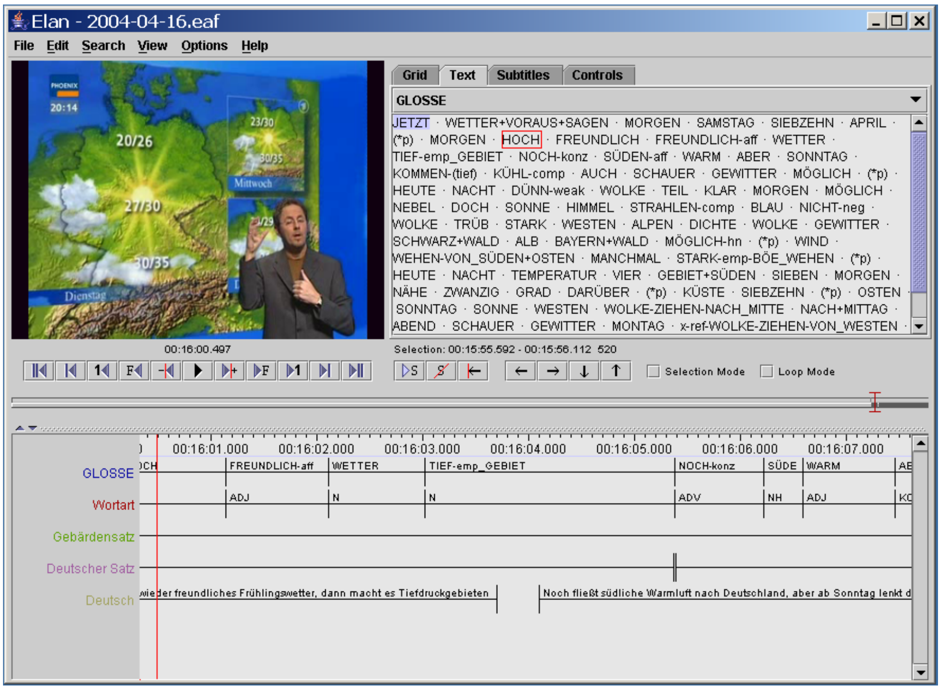
- On the first tier, the signs are annotated as glosses as described above. Every gloss is marked with its starting time and end time.
- While this shows the word boundaries, the DGS sentence boundaries are marked on an additional tier.
- For supporting the alignment of the DGS sentences to the German sentences, a mapping of the DGS to the German sentence boundaries is given too.
- The spoken German sentences with their starting and end timings are given an extra tier.
- For 45 videos parts-of-speech tagging of the DGS glosses is supplied on another tier.
- The last tier gives the information for the locus of the sign in signing space. This is annotated for 20 videos.

The video data is useful for sign language recognition, but the information of the different tiers is useful for supporting the translation too. Additionally all frames of the videos cropped to the window of the interpreter are stored for the recognition system.
In comparison to the ECHO specifications (Nonhebel et al., 2004), our annotation is less detailed, as separate tiers for the dominant hand and for the non-dominant hand or the non-manual sign parameters are not necessary for our task.

5.2. The Bilingual Text-based Corpus

For the sign language translation task, further processing of the video corpus is needed. As we want to provide a text-based corpus, the gloss notation is extracted from the EAF files and stored along with the German sentences as text data.
Long German sentences are split into parts, as these are easier to translate and they resemble an approximation of the DGS sentence length. The correct determination of sentence boundaries in sign languages is still an important issue in linguistic research. However, in our corpus the DGS sentence boundaries are determined according to objective criteria such as e.g. the lowering of the the hands and to the subjective experience of the DGS experts. Furthermore, information that was not translated from German into DGS by the interpreter was deleted.

[3] http://www.mpi.nl/tools/elan.html



GLOSSE

JETZT · WETTER+VORAUS+SAGEN · MORGEN · SAMSTAG · SIEBZEHN · APRIL · (*p) · MORGEN · HOCH · FREUNDLICH · FREUNDLICH-aff · WETTER · TIEF-emp_GEBIET · NOCH-konz · SÜDEN-aff · WARM · ABER · SONNTAG · KOMMEN-(tief) · KÜHL-comp · AUCH · SCHAUER · GEWITTER · MÖGLICH · (*p) · HEUTE · NACHT · DÜNN-weak · WOLKE · TEIL · KLAR · MORGEN · MÖGLICH · NEBEL · DOCH · SONNE · HIMMEL · STRAHLEN-comp · BLAU · NICHT-neg · WOLKE · TRÜB · STARK · WESTEN · ALPEN · DICHTE · WOLKE · GEWITTER · SCHWARZ+WALD · ALB · BAYERN+WALD · MÖGLICH-hn · (*p) · WIND · WEHEN-VON_SÜDEN+OSTEN · MANCHMAL · STARK-emp-BÖE_WEHEN · (*p) · HEUTE · NACHT · TEMPERATUR · VIER · GEBIET+SÜDEN · SIEBEN · MORGEN · NÄHE · ZWANZIG · GRAD · DARÜBER · (*p) · KÜSTE · SIEBZEHN · (*p) · OSTEN · SONNTAG · SONNE · WESTEN · WOLKE-ZIEHEN-NACH_MITTE · NACH+MITTAG · ABEND · SCHAUER · GEWITTER · MONTAG · x-ref-WOLKE-ZIEHEN-VON_WESTEN ·

00:16:00.497

Selection: 00:15:55.592 - 00:15:56.112 520

	00:16:01.000	00:16:02.000	00:16:03.000	00:16:04.000	00:16:05.000	00:16:06.000	00:16:07.000
GLOSSE	)CH · FREUNDLICH-aff · WETTER	TIEF-emp_GEBIET			NOCH-konz	SÜDE · WARM	AE
Wortart	ADJ · N	N			ADV	NH · ADJ	KC
Gebärdensatz							
Deutscher Satz							
Deutsch	wieder freundliches Frühlingswetter, dann macht es Tiefdruckgebieten		Noch fließt südliche Warmluft nach Deutschland, aber ab Sonntag lenkt d				

Figure 2: The ELAN annotation program with five tiers at the bottom: glosses, word classes, sign language sentence boundaries, equivalent German sentence boundaries and the spoken German sentence annotation

For the SMT system the deletion of the non-manual markings assists the translation process. This allows to reduce the number of singletons and the size of the vocabulary. Non-manual components are then added after the translation with post-processing rules again.
Table 1 shows the current state of the bilingual corpus.

	DGS	German
Sentence Pairs	2468	
Number of Running Words	10588	16529
Vocabulary	1895	1302
Singletons	1203	522
Vocabulary without non-manuals	660	1302
Singletons without non-manuals	234	522

Table 1: Corpus statistics of the text-based corpus, where singletons are words occurring only once

Two interesting observations can be made from the corpus statistics:

- First, the number of running words in German is significantly larger compared to DGS. This can be explained by the slower signing compared to spoken utterances in sign language in general. However, (Baker and Padden, 1978) shows that the amount of information transported by the signing is the same as for oral speech. Non-manual markings like the facial expression, the mouth picture or the position and movement of the shoulders contain this additional information.
- As a second observation, the size of the vocabulary in German is about twice the size of the DGS vocabulary without non-manual markings. In addition to the difference in the number of running words, there is another explanation for this phenomenon. The interpreters are working under high pressure, translating the German sentences instantly and without any prior preparations. Therefore they often repeat similar sentences, while in the German part, where the text is written beforehand, several different phrases with the same meaning occur. Some phrases are even poetic in their formulation.

Table 2 gives examples from the bilingual text-based corpus.

5.3. The Monolingual Text-based Corpus

As an additional information source, we supply another text-based corpus, but monolingual. This corpus contains all German weather reports of the Tagesschau from 1999 until today. Transcripts of the news of the last six years are available online[4]. We downloaded all of them and extracted all weather reports.
While the quality of the text is generally good, it does contain several spelling mistakes which explains the rather high amount of singletons. However, the monolingual corpus still allows the improvement of the language model for the sign language translation system. Table 3 gives more details.

6. Conclusion and Outlook

We presented the Phoenix corpus for the language pair German and German Sign Language (DGS) of the domain weather report. It contains annotated video data and processed text, which are suitable for sign language recognition and sign language translation.

[4] http://www.tagesschau.de

DGS	JETZT WETTER+VORAUS+SAGEN MORGEN DONNERSTAG ZWÖLF MAI.
German	Und nun die Wettervorhersage für morgen, Donnerstag, den zwölften Mai.
English	And now the weather forecast for tomorrow, the 12th of May.
DGS	ABER-konz WETTER FREUNDLICH LANG-neg.
German	Das freundliche Wetter ist aber nicht von Dauer.
English	*But the friendly weather is short-lived.*
DGS	WESTEN DEUTSCHLAND SÜDEN TAG REGEN GEWITTER.
German	Im Westen und Süden bilden sich am Tag Schauer und Gewitter.
English	*In the west and south showers and thunderstorms are establishing during the day.*

Table 2: Example sentences of the bilingual text-based corpus for DGS and German; the English translation is provided for the reader

	German Weather Forecast
Sentence Pairs	72724
Running Words	872117
Vocabulary	12320
Singletons	5889

Table 3: Statistics of the German weather report transcriptions

The corpus consists of three parts, first the video data with rich annotations of the signs in glosses, the sentence boundaries and partly with spacial and word class information. Second a bilingual sentence corpus for DGS and German and third a monolingual German weather report corpus.

In addition to the discussed benefits of the Phoenix corpus we want to address some shortcomings too. While interpreted weather reports in detail and interpreted news in general are available every day, the quality of the signing varies depending on the interpreter. Furthermore, the different dialects of the interpreters make the task for both recognition and translation harder, as does the high signing speed. Also, sometimes the incorrect DGS grammar of the interpreters poses a problem too.

Future work will thus include both improvement of the Phoenix corpus and the construction of a new corpus. The Phoenix corpus can be expanded easily by transcribing new recordings which are available daily. Also, more annotation features that hold valuable information should be considered, e.g. the transcription of time lines in DGS. Of course the amount of annotation for word classes and space information can be increased, too.

On the other hand, a new corpus based on video recordings of DGS native speakers can have an improved quality with both better images for recognition and better DGS sentence structures for translation.

7. Acknowledgements

The presented corpus would not have been possible without the dedicated work of Sigrid Kleinen doing the transcription and the help of Lydia Schneider for the additional preparation.

8. References

C. Baker and C. A. Padden. 1978. Focusing on the non-manual components of ASL. In Patricia Siple, editor, *Understanding language through sign language research. (Perspectives in Neurolinguistics and Psycholinguistics)*, pages 27–57, New York, San Francisco, London. Academic Pr.

B. Bauer, S. Nießen, and H. Hienz. 1999. Towards an automatic sign language translation system. In *1st International Workshop on Physicality and Tangibility in Interaction: Towards New Paradigms for Interaction Beyond the Desktop*, Siena, October.

J. Bungeroth and H. Ney. 2004. Statistical sign language translation. In *LREC 2004, Workshop proceedings : Representation and Processing of Sign Languages*, pages 105–108, Lisbon, Portugal, May.

DESIRE. 2004. Aachener Glossenumschrift. Technical report, RWTH Aachen. Übersicht über die Aachener Glossennotation.

J. Heßmann. 2001. *Gehörlos so! Materialien zur Gebärdensprache*. Signum Verlag, Hamburg.

S. Morrissey and A. Way. 2005. An example-based approach to translating sign language. In *Workshop Example-Based Machine Translation (MT X–05)*, pages 109–116, Phuket, Thailand, September.

C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee. 2000. *The Syntax of American Sign Language*. MIT Press, Cambridge, MA, USA.

A. Nonhebel, O. Crasborn, and E. van der Kooij. 2004. Sign language transcription conventions for the echo project. Technical report, Radboud University Nijmegen.

C. B. Traxler. 2000. The Stanford Achievement Test, 9th Edition: National Norming and Performance Standards for Deaf and Hard-of-Hearing Students. *Journal of Deaf Studies and Deaf Education*, 5(4):337–348.

M. Zahedi, D. Keysers, and H. Ney. 2005. Appearance-Based Recognition of Words in American Sign Language. In *IbPRIA 2005, 2nd Iberian Conference on Pattern Recognition and Image Analysis*, pages 511–519, June.