The ATIS Sign Language Corpus

Jan Bungeroth*, Daniel Stein*, Philippe Dreuw*, Hermann Ney*,
Sara Morrissey†, Andy Way†, Lynette van Zijl‡

* Computer Science Department, RWTH Aachen University, 52056 Aachen, Germany
{bungeroth,stein,dreuw,ney}@cs.rwth-aachen.de

† School of Computing, Dublin City University, Ireland
{smorri,away}@computing.dcu.ie

‡ Computer Science Department, University of Stellenbosch, South Africa
lynette@cs.sun.ac.za

Abstract

Systems that automatically process sign language rely on appropriate data. We therefore present the ATIS sign language corpus that is based on the domain of air travel information. It is available for five languages, English, German, Irish sign language, German sign language and South African sign language. The corpus can be used for different tasks like automatic statistical translation and automatic sign language recognition and it allows the specific modelling of spatial references in signing space.

1. Introduction

In order to help deaf people with their everyday communication challenges in a hearing environment, automatic translation systems are developed. However, modern statistical machine translation (SMT) needs to be trained on appropriate data. Unfortunately, language resources for automatic sign language processing are scarce and of varying quality. We therefore introduce a sign language corpus suitable for basic SMT and sign language analysis, the ATIS corpus.

The corpus is based on the Air Travel Information System (ATIS) dataset (Hemphill et al., 1990). It contains transcribed English phrases and sentences of an information service for booking flights and travel information. Of this dataset, 595 sentences were chosen as a base. The corpus was translated with the help of native speakers and is now available for five languages: English, German, Irish sign language (ISL), German sign language (DGS[1]) and South African sign language (SASL).

It is especially interesting for machine translation, as it is limited to one domain. Additionally, it allows the modelling of methods to handle such unique sign language features like the positioning of objects in signing space as it makes extensive use of spatial references refering to airports and other locations.

Furthermore, it could support deaf people for obtaining information in a spoken environment, e.g. at the airport or train station information desk.

2. Related Work

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

- The European Cultural Heritage Online organization (ECHO)[2] published corpora for Swedish sign language, British sign language and the sign language of the Netherlands. These corpora contain children's fables and poetry each signed by a single signer. However, they have a large vocabulary which makes automatic learning difficult.

- 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[3] 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.

- The Signs of Ireland corpus developed at the Centre for Deaf Studies, Dublin (Leeson et al., 2006) contains video data of approximately 40 Deaf ISL users collected over 3 years. Participants tell personal narrative, a children's story and sign elicited sentences. The corpus is hand annotated.

- (Y.-H. Chiu and Cheng, 2007) perform SMT experiments on a corpus of about 2000 sentences for the language pair Chinese and Taiwanese sign language.

- For the domain weather reports, a corpus of 2468 sentences in German and DGS was reported by (Bungeroth et al., 2006). It is particularly used for SMT and sign language recognition.

- The RWTH-BOSTON-104 Database (Dreuw et al., 2007) contains 201 sentences in American sign language with English annotations. This corpus is mainly used for automatic sign language recognition.

3. Corpus Setup and Notation

The sentences from the original ATIS corpus are given in written English as a transcription of the spoken sentences.

[1] Deutsche Gebärdensprache

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

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

The domain is restricted to flight information and booking services. Table 1 gives several example sentences from the original ATIS corpus that were selected for the ATIS sign language corpus. In total 595 sentences were chosen for the sign language translation.

what flights depart on Friday
leaving Sunday after twelve noon
please list the earliest lunch flight from Dublin to London
flights from Liverpool to Dublin arriving before five p.m.
show me the fares from Dublin to London

Table 1: Example sentences from the original ATIS corpus

3.1. 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 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. Basically, in this work the specifications of the Aachener Glossenumschrift (DESIRE, 2004) are followed in this work. However, due to different groups that worked on the different translations, variations in the gloss notation style occur.

As an example, the sentences in Table 2 in English and SASL are taken from the ATIS corpus.

Here the `_a` and `_b` denote locations in signing space, representing the cities of Dublin and Cork in the first example. The `-qu` represents an interrogative non-manual expression of the face. In the second example, the `-X_aerlingus` is used as a reference in signing space of the noun `FLIGHT`, which has the special meaning of the specific air carrier in this case. The ++ is the superlative of the adjective `LATE`.

3.2. The ISL Corpus

The ISL corpus formed the first translation into SL of the ATIS data. Two Deaf native ISL signers were engaged to assist in accurate translation. They worked in tandem translating, signing and monitoring as each sentence was captured on video. Some alterations, such as changing place names to local names, were made in order to facilitate signing. ELAN Annotation Software[4] was used to facilitate gloss annotation of the ISL data. The ELAN based annotation also provides precise starting and stopping times for each sign that is annotated as gloss. This information is valuable for automatic sign language recognition because a sign shown in a video segment can be related to its semantic representation (i.e. its gloss).

3.3. The DGS Corpus

All sentences were first translated into German. A deaf native DGS speaker and two bilingual experts used the German sentences as reference for the translation into DGS. The special grammatical features of DGS were given special attention. Thus, the DGS translation heavily utilizes the signing space for refering to the different airport locations that are seen quite often in the corpus.

3.4. The SASL Corpus

The basis for the translation into SASL were the English sentences. Twenty sentences covering all different aspects of the corpus were chosen and carefully translated by a deaf native SASL speaker. This was recorded and annotated into glosses. From this basis, the remaining sentences were translated into glosses too. Again, emphasis was put on capturing the location placement in signing space correctly. Figure 1 shows an image frame from the SASL recordings where the signer signs `FLIGHT-X`, pointing (`X`) with his dominant hand to the sign `FLIGHT` shown by his non-dominant hand.

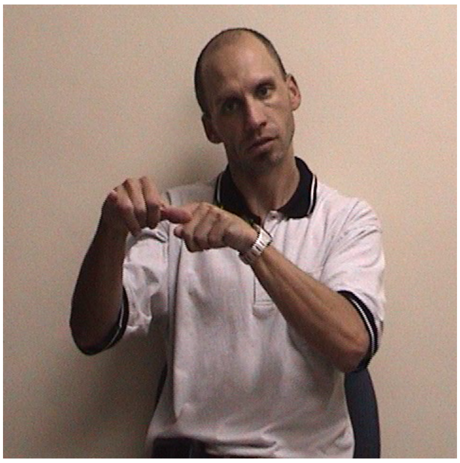

Figure 1: Image frame taken from an SASL video showing the sign `FLIGHT-X`

3.5. Spatial References

A specific feature of the ATIS sign language corpus is the usage of spatial references that refer to objects placed in signing space. Figure 2 shows the SASL sentence DUBLIN_a a_FLY_b CORK_b with its representation in signing space. Here location a (DUBLIN) is placed on the right side of the signer and location b (CORK) to the left side. Between the locations the verb FLY is signed in an arc shaped movement from a to b.

These spatial references that occur in all sign language translations of the ATIS corpus need to be addressed specifically by translation and recognition algorithms, e.g. the placement of objects have to be modelled accordingly in the translation process.

3.6. Corpus Statistics

As already mentioned, the ATIS corpus contains 595 sentences in five languages. Using the corpus for SMT systems

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

DUBLIN_a a_FLY_b CORK_b PRICE WHAT-qu
What is the price of flights from Dublin to Cork?
FLIGHT-X_aerlingus LATE++ FLIGHT
Which is the latest flight that Aer Lingus has?

Table 2: Example SASL sentences showing the gloss notation

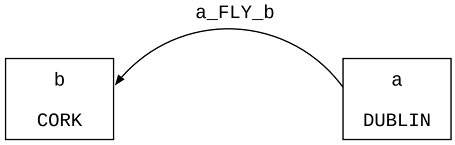

Figure 2: SASL sentence "DUBLIN_a a_FLY_b CORK_b" showing the locations in signing space

requires a splitting into a training set, a development set and a test set. The training set is used for the learning process of the system, where repeated phrases are memorized. The development set helps to optimize the parameters of the system and the test set is used for evaluation only. Table 3 shows a detailed breakdown of the sets. Here, singletons are words occurring only once within a set.

Table 4 gives several examples of sentences in all languages. Note the different annotation styles used by the different translation team members.

4. Experimental Results

The ATIS corpus was used in a number of experiments for automatic statistical machine translation and automatic sign language recognition.

For the translation part, experiments were reported on the corpus by (Morrissey et al., 2007) and (Morrissey, 2008) for the language pairs ISL–English, ISL–German, DGS–English and DGS–German. First experiments on the language pairs SASL–German and SASL–English show similar baseline results as for the already reported ones. This provides feedback information of the continuous quality of the corpus.

In (Stein et al., 2007) the corpus is additionaly used in the context of a data-driven sign-language-to-speech system. Here, automatic sign language recognition was applied to the ISL videos. This is a more demanding task though as the vocabulary size is larger than in similar corpora used for recognition.

5. Conclusion

We introduced the ATIS corpus that aims at statistical machine translation, automatic sign language recognition and further natural language processing for the languages English, German, Irish sign language, German sign language and South African sign language. We explained the notation system used and how the corpus was assembled for the different sign languages. We show that the corpus was already used by a number of experiments that led to publications in the mentioned fields of research.

Future work will focus on conducting further experiments using different approaches in machine translation to cope with the scarce data problem. Further translations to other sign languages are possible too.

6. Acknowledgements

We want to thank Kevin G. Mulqueen and Mary Duggan for their assistance in the creation of the ISL corpus. Also, we express our gratitude to Bennie Botha for helping with the SASL translation and for his patience with the video recordings. Finally we also want to thank Ege Karar and Horst Sieprath for their helpful input on the details of the German sign language.

7. 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, May.

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

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney. 2007. Speech recognition techniques for a sign language recognition system. In *Interspeech*, pages 2513–2516, Antwerp, Belgium, August.

C.T. Hemphill, J.J. Godfrey, and G.R. Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 96–101, Hidden Valley, PA., June.

L. Leeson, J. Saeed, A. Macduff, D. Byrne-Dunne, and C. Leonard. 2006. Moving Heads and Moving Hands: Developing a Digital Corpus of Irish Sign Language. In *Proceedings of Information Technology and Telecommunications Conference 2006*, Carlow, Ireland.

S. Morrissey, A. Way, D. Stein, J. Bungeroth, and H. Ney. 2007. Towards a hybrid data-driven mt system for sign languages. In *Proc. of the 11th Machine Translation Summit*, pages 329–335, Skövde, Sweden, September.

S. Morrissey. 2008. *An Exploration of Data-driven Machine Translation for Sign Languages*. Ph.D. thesis, Dublin City University, Dublin.

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

D. Stein, P. Dreuw, H. Ney, S. Morrissey, and A. Way. 2007. Hand in hand: Automatic sign language to speech translation. In *11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 214–220, Skövde, Sweden, September.

		EN	DE	ISL	DGS	SASL
Train	no. sentences	418				
	no. running words	3008	3544	3028	2980	1825
	vocab. size	292	327	265	244	215
	no. singletons	97	118	71	84	76
Dev	no. sentences	59				
	no. running words	429	503	431	434	233
	vocab. size	134	142	131	119	96
Test	no. sentences	118				
	no. running words	999	856	874	877	477
	vocab. size	174	158	148	135	111

Table 3: Corpus Overview

Language	Sentence
English	What flights are there from Belfast to Dublin?
German	Welche Flüge gibt es von Belfast nach Dublin?
ISL	WHAT FLIGHT be-BELFAST FROM BELFAST TO DUBLIN be-DUBLIN
DGS	BELFAST IX_a a_BIS_b DUBLIN IX_b a_FLIEGEN_b WAS-qu
SASL	BELFAST_a a_FLY_b DUBLIN_b X WHICH-qu
English	Which flights arrive in Dublin at or before eight p.m. on Friday?
German	Welche Flüge kommen in Dublin vor acht Uhr Abend am Freitag an?
ISL	WHICH FLIGHTS DUBLIN be-DUBLIN FLY ARRIVE FLY ARRIVE EIGHT OR BEFORE EIGHT ON FRIDAY
DGS	FREITAG ACHT UHR ABEND DUBLIN IX_a FLIEGEN_a WAS-qu
SASL	FRIDAY FLIGHT-X TOUCH-DOWN DUBLIN AROUND EIGHT-OCLOCK WHAT-qu
English	Hi can I get a one way ticket from Cork to Dublin?
German	Hi, kann ich ein Ticket für den Hinweg von Cork nach Dublin bekommen?
ISL	HI CAN ONE TICKET ONE WAY CORK TO DUBLIN
DGS	HALLO CORK IX_a a_BIS_b DUBLIN IX_b a_FLIEGEN_b TICKET MÖGEN
SASL	HELLO CORK_a a_FLY_b DUBLIN_b THAT'S-ALL
English	Show me all flights that depart before ten a.m. and have first class.
German	Zeigen Sie mir alle Flüge die vor zehn Uhr morgens abfliegen und erste Klasse haben.
ISL	SEE ALL FLIGHT BEFORE TEN O'CLOCK AND FIRST CLASS
DGS	ZEIGEN_self FLIEGEN ++ ABFLUG VOR ZEHN UHR FRÜH UND ERSTE KLASSE DABEI
SASL	SHOW ALL FLIGHT-X FIRST CLASS TAKE-OFF BEFORE TEN-OCLOCK
English	Cheapest fare one way.
German	Günstigster Flugpreis für Hinflug.
ISL	CHEAP ONE WAY
DGS	TARIF BILLIG-emp a_FLIEGEN-emp-nostop_b
SASL	a_FLY_b THAT'S-ALL CHEAP++ PRICE

Table 4: Several example sentences taken from the ATIS sign language corpus in five languages

H.-Y. Su Y.-H. Chiu, C.-H. Wu and C.-J. Cheng. 2007. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29, no. 1:28–39.

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.