Morpho-Syntax Based Statistical Methods for Automatic Sign Language Translation

Daniel Stein, Jan Bungeroth and Hermann Ney
Lehrstuhl für Informatik VI,
Computer Science Department, RWTH Aachen University
D-52056 Aachen, Germany
{stein,bungeroth,ney}@i6.informatik.rwth-aachen.de

Abstract

We present a novel approach for the automatic translation of written text into sign language. A new corpus focussing on the weather report domain for the language pair German and German Sign Language is introduced. We apply phrase-based statistical machine translation, enhanced by pre- and post-processing steps based on the morpho-syntactical analysis of German. Detailed results are given based on automatic and manual evaluation.

1 Introduction

The aim of this work is to employ an automatic translation system from written German to German Sign Language (DGS[1]), the primary means of communication for the deaf people in Germany.

It may seem surprising at first to propose a translation of written text if the target group typically has no visual impairment. However, (Traxler, 2000) shows that the majority of the deaf community possesses only poor to moderate reading skills. The lack of auditory feedback and the still common practice of oral teaching are two responsable factors.

In this paper, we present our translation system from German sentences into DGS, a language independent from German. The results are visualized using an avatar developed for the presentation of sign languages (Elliott, Glauert, Kennaway, & Marshall, 2000). We investigate in how far morpho-syntactic pre- and post-processing can enhance the translation results. We also investigate the specific demands for sign language translation. Finally, we present detailed results based on both automatic and manual evaluation of the translation output.

[1] Deutsche Gebärdensprache

1.1 State-of-the-Art

Several researchers deal with the challenges of automatic sign language translation. To the best of our knowledge, statistical methods have not been used in translation from written text to sign language yet. Moreover, the other approaches did not present quantitative results. Thus, performance comparison is not possible.

(Morrissey & Way, 2005) investigate corpus-based methods for example-based sign language translation from English to the sign language of the Netherlands. With the small corpus and no available lexicon, the system is robust for sentences already encountered in the training set, but has problems with unseen combinations of corpus chunks as well as corpus parts that it is unable to align.

(Huenerfauth, 2004) explores the challenges of machine translation techniques from written text to sign languages. He proposed methods concerning the theoretical issues arising during translation, for example a notation for signs which use the 3D space around the signer to form complex expressions.

(Sáfár & Marshall, 2001) propose a decompensation of the translation process into two steps: first they translate from written text into a semantic representation of the signs. Afterwards a second translation into a graphically oriented representation is done. Both steps use rule-based techniques for a specific domain in British Sign Language. However, no quantitative results were published.

(Bauer, Nießen, & Hienz, 1999) propose the recognition of captured sign language videos into manual sign parameters. They argue that these parameters can be transformed into written text by statistical machine translation. However, no detailed results are given.

2 Phrase-Based Machine Translation

We use a statistical machine translation system to automatically transfer the meaning of a source language sentence into a target language sentence (Zens et al., 2005). Following the notation convention, we denote the source language with J words as $f_1^J = f_1 \dots f_J$, a target language sentence as $e_1^I = e_1 \dots e_I$ and their correspondence as the a-posteriori probability $\Pr(e_1^I|f_1^J)$. The sentence $\hat{e}_1^I$ that maximizes this probability is chosen as the translation sentence as shown in Equation 1.

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \left\{ \Pr(e_1^I|f_1^J) \right\} \quad (1)$$

$$= \operatorname*{argmax}_{e_1^I} \left\{ \Pr(e_1^I) \cdot \Pr(f_1^J|e_1^J) \right\} \quad (2)$$

The estimation of the a-posteriori probability is divided into three subproblems:

1. the language model, for which we employ trigrams smoothed with Kneser-Ney discounting (Chen & Goodman, 1998)

2. the translation model, where we use the phrase-based translation as described in (Zens et al., 2005)

3. the search algorithm finding the best path. We use monotone search and reordering constraints (Kanthak, Vilar, Matusov, Zens, & Ney, 2005), which are explained in the next section

2.1 Reordering constraints

Closely related language pairs, for example Catalan-Spanish, have a very similar grammar structure, so that their phrases have the same sequence over large portions of the text. For the search algorithm looking for the best translation, the search space can be reduced if we assume monotone word dependency. However, many other language pairs differ significantly in their word order. To keep computational costs at a reasonable scale, we allow a larger search space but limit the permutation number by *reordering constraints.*

A reordering constraint is a directed, acyclic graph that allows limited word reordering of the source sentence. The edges of each possible path equal a permutation π of the numbers 1 to J.

In our work, we investigate the influence of three reordering graphs (Figure 2)(Kanthak et al., 2005) on our translation results: the local constraint, the IBM constraint and the inverse IBM constraint. Each graph allows characteristic permutation types, constrained by a window size w: the local constraint allows each word in the sentence to be moved up to a maximum of $w-1$ steps towards the front or the end of the sentence. The IBM constraint allows up to $w-1$ words in the sentence to be moved to the end of the sentence, likewise, the inverse IBM constraint allows up to $w-1$ words to be moved to the sentence beginning.

The higher the window size w, the higher the amount of possible permutations has to be considered. A window size which is higher or equal to the sentence length J results in a search space that is equal to the maximum of permutations possible.

2.2 Evaluation Criteria

In our experiments, we use the following criteria for evaluating the translation results:

Word Error Rate (WER): The WER is computed as the edit distance between the produced translation and the reference translation based on the Levenshtein alignment (i.e. the minimum number of required insertions, substitutions and deletions to match the two sentences).

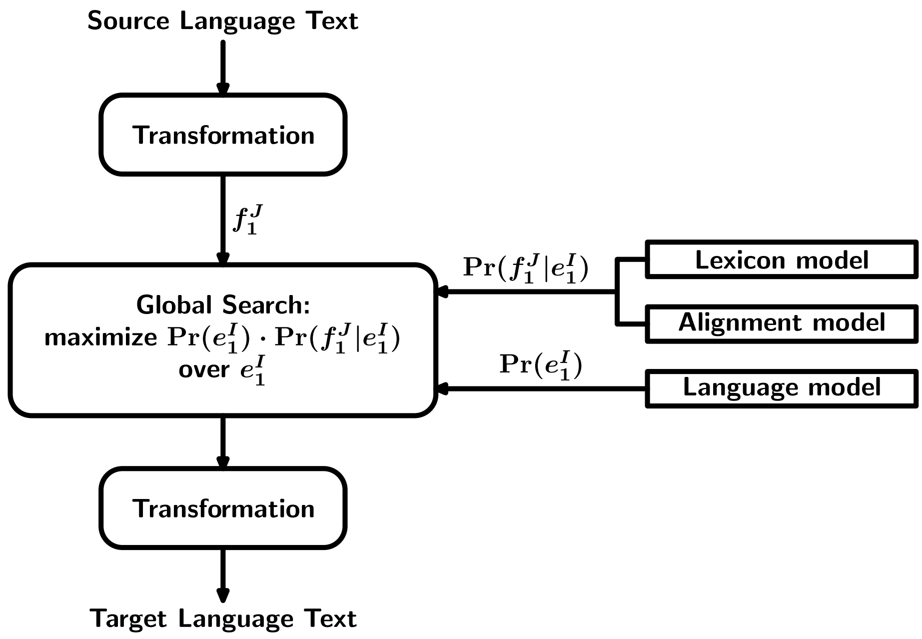


Figure 1: Architecture of the translation approach based on Bayes' decision rule

Position-independent Word Error Rate (PER): To overcome the problem of a possibly misleading WER due to the dependency on the perfect word order, we introduce the PER as an additional measure which ignores the order of the words when comparing the words of the produced translation and the reference translation.

3 Sign Language

Language research has long been tied exclusively to spoken languages. Only about forty years ago the first serious investigations of sign languages have begun. Sign languages are communication systems which have evolved over generations of deaf signers and are not derived from spoken language. Like all natural languages, no international sign language exists, and even DGS has several dialects.

Grammar and vocabulary differ from the ones used in spoken language. Moreover, the unique possibilities of a visual-gestural based language allow a specific grammar which employs the usage of space and facial expressions to bestow additional language- and meta-language information (Braem, 1995). In DGS, no articles are used and no copula can be found. DGS also makes extensive use of the spatial feature to flex and derive its words.

For example, time information is conveyed by a spatial principle. In most western European sign languages, imaginary lines can be found that represent different moments in time. These so-called time-lines starts from the back of the body, which represents things past tense. It then moves on to the middle layer which extends itself just to the front of the signer, or present tense. Signs executed to the front in some distance of the signing person are usually related to the future. Thus, the signs for 'TOMMOROW' are placed on the front end of this time line, while signs like 'YESTERDAY' can be found on its back end.

While it would go beyond the scope of this work to give a complete overview of the different grammar characteristics of DGS, some selected phenomena will be presented here.

3.1 Verb Flexions

DGS belongs to the group of languages where the word flexion is more important than the word position in the sentence. Flexed verbs usually share the same root, which means that they are mostly identical in its components, but differ in such

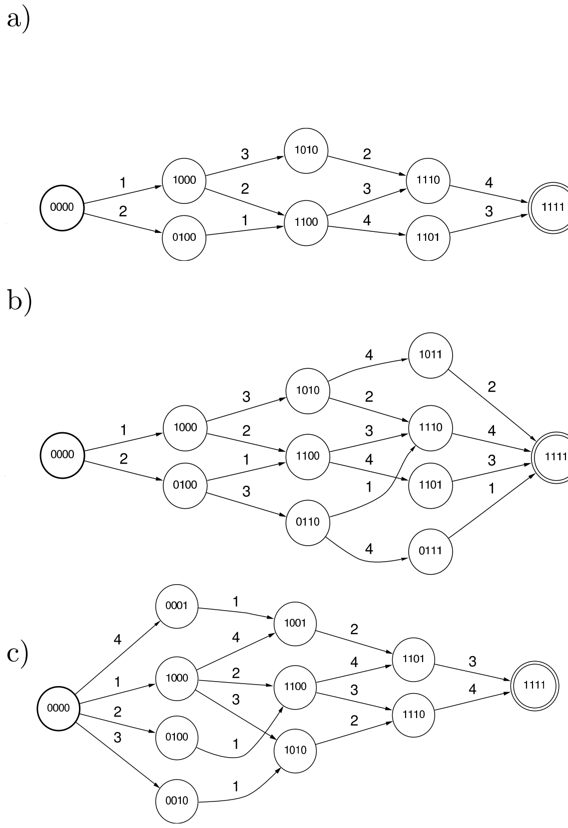


Figure 2: Permutation graph of a source sentence $f_1 f_2 f_3 f_4$ using a window size $w = 2$ for a) local constraints, b) IBM constraints and c) inverse IBM constraints

elements as movement speed, direction or amount of signing space used.

For undefined pronouns not present in a conversation, the direction of a verb indicates subject and object and number of occurrences. For this, a predefined set of movements is used to differ between casus and numerus. For example, in Figure 3 (Braem, 1995) the root verb 'to give' is flexed in space to indicate that this verb is flexed in 2nd person plural 'I give you all'.

Another group of verbs is coordinating the directions and locations with specific persons previously 'stored' in locations within the signing space. This storage is done by executing the sign in a specific place and thus 'storing' it there for later references (Wrobel, 2001). By flexing a verb towards this so-called discourse entity, the signer is referring back to this person like in a pronoun ('She is giving him an apple'). Normally, the starting point is referencing to the subject and the end point to the direct or indirect object. This technique is called *verbal agreement.*

Since every location is clearly defined in regards of what person it references, verbal agreement is in some ways more exact than verbal flexion in some vocal languages. Locations can also be used to reference objects, abstract concepts or sentences.

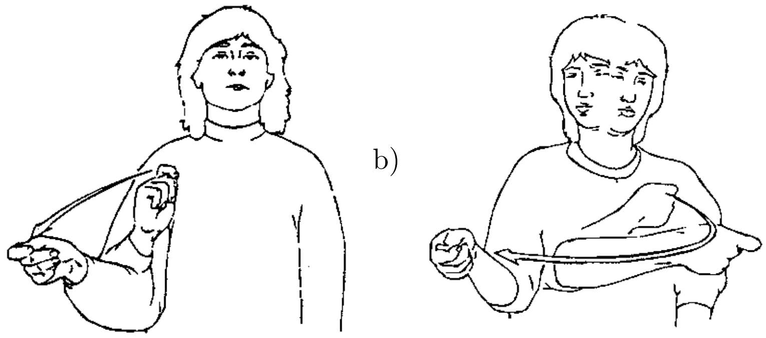
3.2 Incorporation

Another important feature of sign languages is their ability for *incorporation.* This means that they are able to deliver parallel information on different sub-levels. For example, additional information can be delivered through the non-manual devices. Facial expression and head position can be used to indicate questions, negations and sub-clauses (for American Sign Language, see (Sandler, 1999)); the upper part of the body can be turned to indicate a role change of the speaking character in direct speech, and the lips can be used to discriminate between signs which have the same manual components, specificate subordinated signs or carry other additional information. Conditional subclauses are indicated in DGS through raised eyebrows and a slight tilt of the head. Non-manuals are also used in some adjectives and adverbs to modify signs with what would be 'intensive', 'big' or 'uncomprehending' in German.

But even without non-manual devices, a sign can already incorporate many informations. For each word in the term 'in three days' different signs exist, yet they are merged to a new, single sign which contains aspects of all of the three separate signs but translates the whole term.

Incorporation can be found on prosodic, semantic and syntactic levels. Because of the simultaneous usage of these features, DGS is often referred to as being a parallel language, compared to a sequential language such as German.

For more information about the grammar of DGS, the reader is referred to (Braem, 1995) and (Prillwitz, 1985).



a) b)

Figure 3: Example for verb flexion in DGS: a) 'to give' b) 'I give you all'

3.3 Notation System

For DGS, there is no official written form. Depending on transcription purpose, existing systems differ in accuracy and detail depth.

Stokoe argued in his work (Stokoe, Casterline, & Croneberg, 1965) that there are three aspects of manual sign articulation, namely hand configuration, place of articulation and movement. This model was extended by the hand situation as a fourth parameter (Klima & Bellugi, 1979)(Battison, 1978).

The validity of these parameters can be seen in several observations: for every one of these components, *minimal pairs* can be found in DGS, i.e. two semantically distinct words in a language which differ only in one particular component (Figure 4) (Braem, 1995).

Moreover, similar phenomena to the slip of the tongue have been reported on deaf signers (Howard Poizner & Bellugi, 1990). These so-called slips of the hand, that is the misconception of single signs that still follow allowed sign production rules, sometimes only differ in only one of these parameters.

The four sub-lexical manual components form the basis for several notation systems, including HamNoSys, to be explained below. They also have been tested as a starting point for sign language recognition using hidden Markov models in (Bauer & Kraiss, 2001).

In addition to these manual components, other features need to be recorded as well. People not accustomed to sign languages tend to think of hands and arms only, whereas the non-manual devices, which include facial expression, eye gaze, head tilt and body posture, carry meta-language information only. However, one of the first studies conducted with non-manual devices showed that it is still possible to gain some information about the conversation topic if the hands of the signers cannot be seen (Baker & Padden, 1978), (Corina, Bellugi, & Reilly, 1999). Thus, both manual and non-manual components are vital for signing.

For our work, we use so-called glosses, a semantic representation of the sign language. As a convention, the meaning of the sign is written as the upper case stem form of the corresponding word in a spoken language. Our gloss notation is a variety of the Aachener Glossenumschrift, developed and maintenanced by the Deaf Sign Language Research Team (DESIRE), Aachen. For our translation, it annotates all important sign language grammar features.

The following example can be translated into English with 'The high pressure areas over the atlantic ocean are growing larger'.

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

The three signs are transcripted 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.

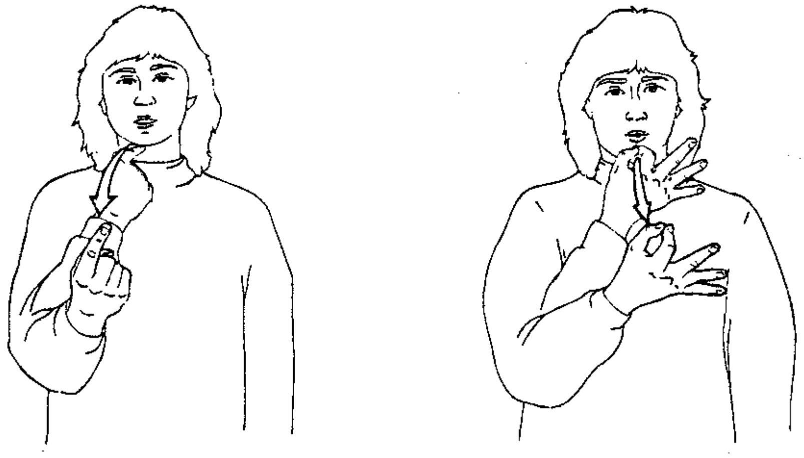


Figure 4: Minimal pair for the hand configuration in DGS: 'SAY' and 'ASK'

4 Experiments

The corpus used in this work was manually transcribed by language experts. On the German television channel Phoenix, the German weather forecast is translated into DGS. The videos, i.e. the German sentences spoken by the announcer and the signs from the interpreter, were transcribed, and their quality were checked on a regular basis.

The corpus statistics are listed in Table 1. For a detailed description, the reader is referred to (Bungeroth, Stein, Dreuw, & Zahedi, 2006).

4.1 Morpho-Syntax Based Pre-Processing

We try to enhance the translation by either omitting redundant sentence information or by transforming parts that do not change the meaning of the sentence in the pre-processing phase. These measurements are especially important on smaller corpora.

In our work, we employ the gerCG parser[2] for various pre-processing steps. gerCG delivers all vital parts-of-speech information (POS). We also employed a parser that reads the gerCG tags as well a a simple rule file which lists actions for the specific POS. In informal experiments on the development corpus, several rule files with different actions were compared against each other.

[2] `http://www.lingsoft.fi`

Among the procedures tested were the transformation of nouns into stem form. While DGS is a highly flexed language, gender information provided in the affixes of the German words are not translated and can be thus omitted. Also, if the relation between subject and object is usually apparent, the words can be reduced to stem form completely.

Another pre-processing step that leads to improvement is the splitting of words at break points. German grammar allows concatenation of small words to form a compound word often unseen in training data. However, since DGS works with compounds, too, trying to translate the single parts of the compound part should improve the error rates.

As a third pre-processing step, German POS commonly not used in DGS were deleted. Among them were mainly articles and certain conjunctions.

4.2 Post-Processing

Post-processing tries to circumvent typical errors of the translation algorithm.

Difficult to handle in DGS translation are so-called discourse entities – stored persons, names or even sentences – which can be referenced to by pointing at them or executing a sign using this specific area. We marked the position of all entities appearing in our corpus. Signs that are usually as-

Table 1: Statistics of the Phoenix corpus

		DGS	German
training set	sentences	2272	
	number of running words	9947	15124
	vocabulary size	640	1246
	number of singletons	223	504
development set	sentences	98	
	number of running words	496	736
	vocabulary size	183	274
	number of singletons	13	24
test set	sentences	98	
	number of running words	486	732
	vocabulary size	184	304
	number of singletons	9	35

signed to a specific position, for example for geographical reasons, have been annotated in a database.

Emphasis and comparative degree that share the same stem word in both languages were treated as stem form during training and translation. The deleted information was added in the post-processing step. Note that this step has no effects on the automatic error rates WER and PER, but will only influence the human evaluation.

4.3 Results

We investigated in how far the grammar transformations influence the error rates. To avoid training on testing, all optimizations have been conducted on the separate development set, optimizing on the PER. All results use a reference file with 2 correct translations average. Discarding not-needed POS in German already improves the result greatly. It seems that many words occuring in German disorient the algorithms since they are not needed in DGS. Stem forms reduce the vocabulary size and also the number of out-of-vocabulary words. Splitting the nouns helps to enhance the translation quality, too, since unknown German word compounds are fragmented into smaller word parts. The results are listed in Table 2. In total, we improve the baseline by 9.0% in the WER and by 9.6% in the PER.

If we employ constraints, then the best result is achieved for local reordering and a window size of 2: the PER improves to 27.4 and the WER goes slightly up to a WER of 38.2 (Table 3). While we expected more enhancement from this approach, the translations in the corpus are made by hearing interpreters under extreme time pressure conditions. We argue that their grammar might be too close to the German grammar for the reordering constraints to work properly.

For human evaluation, we asked two human experts (both congenitaly deaf) to rate the coherence of a German sentence to the avatar output with numbers ranging from 1 (uncomprehensive) to 5 (perfect match). For this purpose, we took the first 30 sentences from the test corpus and evaluated both the reference sentences output and the

Table 2: Results of a concatenation of the preprocessing steps, measured on the development corpus

	WER	PER
baseline	48	37.8
+ discard conj. and articles	40.4	30.0
+ stem form	39.2	29.8
+ split nouns	37.0	28.2

Table 3: Results on the Phoenix corpus

	WER	PER
baseline	48.0	37.8
best result	**38.2**	**27.4**

Table 4: Translation examples

	reference sentence	translated sentence
correct	JETZT WETTER+VORAUS+SAGEN MORGEN SAMSTAG ZWÖLF MÄRZ	
equivalent	AUCH NORDEN+WESTEN BEREICH_nordwesten WOLKE REGEN ZIEHEN_nach_südosten	NORDEN+WESTEN BEREICH_nordwesten AUCH WOLKE REGEN ZIEHEN_von_nordwesten
wrong	TIEF FRANKREICH ZIEHEN_nach_frankreich	TIEF ZIEHEN_nach_mitte

Table 5: Human evaluation results

human expert A	average score
reference	3.3
translation	2.9
human expert B	average score
reference	3.7
translation	3.4

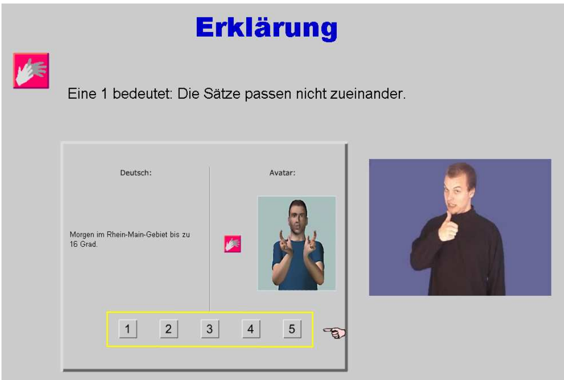


Figure 5: Evaluation tool for comparison of German sentence and avatar output. Introduction is translated simultaneous with pre-recorded movie clips.

translation sentences output of the avatar, for a total of 60 sentences.

The rating difference of reference and translation sentences is 0.4 average points for the first expert and 0.3 points for the second. The results in general are still low (i.e., at 3.3 average). The focus of this work was on the implementation of the translation algorithms and the avatar was only supported poorly, however, it seems that the results for human evaluation and the results for automatic evaluation are comparable.

4.4 Translation Examples

In Table 4 some examples for translation are given. The first example is fairly easy and close to the German grammar. The translation made no mistakes. The sentences of the second example are semantically equal but differ in synonyms and word sequence. The translation gets an error penalty both in WER and PER, but not in the human evaluation. In the last example, the translation algorithm did not know the word 'Frankreich' and omitted it in the translation. Therefore, the sentence makes no sense anymore.

5 Conclusion

In this paper, we present the first phrase-based statistical machine translation approach for a sign language. A new corpus based on bilingual weather reports is introduced. We showed how a morpho-syntactic knowledge source for German can be used to significantly improve the translation quality. For this, we came up with a flexible POS parser that allowed us to transform the words according to linguistic assumptions. The results of the different methods have been compared against each other.

For important features of DGS which are hard to translate statistically such as incorporation and space information we implemented pre- and post-processing methods.

References

Baker, C., & Padden, C. A. (1978). Focusing on the nonmanual components of ASL. In P. Siple (Ed.), *Understanding language through sign language research. (perspectives in neurolinguistics and psycholinguistics)* (p. 27-57). New York, San Francisco, London: Academic Press.

Battison, R. (1978). *Lexical borrowing in american sign language.* Silver Spring, MD: Linstok Press.

Bauer, B., & Kraiss, K.-F. (2001, April). *Towards an automatic sign language* *recognition system using subunits.* London: Springer.

Bauer, B., Nießen, S., & Hienz, H. (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.

Braem, P. B. (1995). *Einführung in die gebärdensprache und ihre erforschung* (Third ed.). Hamburg, Germany: Signum-Verlag.

Bungeroth, J., Stein, D., Dreuw, P., & Zahedi, M. (2006). A german sign language corpus of the domain weather report. In *5th international conference on language resources and evaluation (to appear).* Genoa, Italy.

Chen, S. F., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling* (Tech. Rep. # TR-10-98). Cambridge, MA: Harvard University.

Corina, D. P., Bellugi, U., & Reilly, J. (1999). Neuropsychological studies of linguistic and affective facial expressions in deaf signers. *Language and Speach*, *42*, 307-331.

Elliott, R., Glauert, J. R. W., Kennaway, J. R., & Marshall, I. (2000). The development of language processing support for the visicast project. In *Assets '00: Fourth international acm conference on assistive technologies* (pp. 101–108). New York, NY, USA: ACM Press.

Howard Poizner, E. S. K., & Bellugi, U. (1990). *What the hands reveal about the brain* (Fifth ed.). MIT Press.

Huenerfauth, M. (2004). A multi-path architecture for machine translation of English text into American Sign language animation. In *Student workshop at human language technolgies conference hlt-naacl.* Boston, MA, USA.

Kanthak, S., Vilar, D., Matusov, E., Zens, R., & Ney, H. (2005). Novel reordering approaches in phrase-based statistical machine translation. In *Acl workshop on building and using parallel texts* (pp. 167–174). Ann Arbor, MI: Association for Computational Linguistics.

Klima, E. S., & Bellugi, U. (1979). *The signs of language.* Cambridge, UK: Harvard University Press.

Morrissey, S., & Way, A. (2005). An example-based approach to translating sign language. In *Workshop example-based machine translation (mt x–05)* (p. 109-116). Phuket, Thailand.

Prillwitz, S. (1985). *Skizzen zu einer Grammatik der Deutschen Gebärdensprache.* Hamburg, Germany: Signum Verlag.

Sáfár, É., & Marshall, I. (2001). The Architecture of an English-Text-to-Sign-Languages Translation System. In G. A. et al (Ed.), *Recent advances in natural language processing (ranlp)* (p. 223-228). Tzigov Chark, Bulgaria.

Sandler, W. (1999). Prosody in two natural language modalities. *Language and Speech*, *42*, 127-142.

Stokoe, W., Casterline, D., & Croneberg, C. (1965). *A dictionary of american sign language on linguistic principles.* Washington D.C., USA: Gallaudet College Press.

Traxler, C. B. (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.

Wrobel, U. R. (2001). Referenz in gebärdensprachen: Raum und person. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, *37*, 25-50.

Zens, R., Bender, O., Hasan, S., Khadivi, S., Matusov, E., Xu, J., Zhang, Y., & Ney, H. (2005). The RWTH phrase-based statistical machine translation system. In *Proceedings of the international workshop on spoken language translation (iwslt).* Pittsburgh, PA, USA.