Handling Multimodality and Scarce Resources in Sign Language Machine Translation

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Informatiker
Christoph Andreas Schmidt
aus Mainz

Berichter: Professor Dr.-Ing. Hermann Ney
Professor Dr. Ruben San-Segundo Hernández


Diese Dissertation ist auf den Internetseiten der Hochschulbibliothek online verfügbar.
Hiermit versichere ich, dass ich die vorliegende Doktorarbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel verwendet habe. Alle Textauszüge und Grafiken, die sinngemäß oder wörtlich aus veröffentlichten Schriften entnommen wurden, sind durch Referenzen gekennzeichnet.

Bonn, im Sommer 2016

Dipl.-Inform. Christoph Andreas Schmidt
Abstract

In the field of statistical machine translation, the translation of sign languages poses an interesting and challenging problem. Translating from a signed language into a spoken language is usually a two-step process. From the video of a person signing, an automatic sign language recognition system extracts the signs in some form. Since signed languages differ in grammar, vocabulary and expression from spoken languages even within the same country, the recognised sequence of signs has to be translated into a spoken language text.

In sign languages, meaning is conveyed simultaneously not only via the two hands, but also by facial expressions, body posture, head movement, and eye gaze. Because of this complex and multimodal nature of sign languages, there is no common writing system, and the scientific question of an annotation scheme suitable for machine translation remains open. Another difficulty when applying statistical methods to sign language translation is the lack of a sufficient amount of training data. This data scarcity often leads to poor translation results. Moreover, the multimodal nature of sign language is not handled by current translation systems, which usually process sequences of words.

In this thesis, we approach the above three problems: finding a suitable annotation scheme, dealing with small amounts of annotated data, and handling multimodality in the machine translation process. While we focus on the second step of translating the recognised signs into a spoken language text, we also aim at improving the overall process of recognition and translation by optimising the interface between the two systems. In the course of the EU-project SignSpeak, our research group implemented the whole pipeline of sign language recognition and translation, and we evaluated this pipeline with automatic measures and human evaluators.

The annotation scheme of the RWTH-PHOENIX-Weather 2014, a sign language corpus which was created in the course of this thesis, contains not only information on the signs expressed by the hands, but also marks mouthings, locations in the signing space or simultaneous signing of two different signs with both hands. We analyse the importance of this additional information for machine translation and devise an improved way of including it in the process of translation.

Since available sign language corpora are rather small when compared to spoken language corpora, the lack of sufficient training data often leads to a poor automatic alignment between the annotated signs and their translations in the spoken language. This issue is aggravated by the fact that sign languages, which are expressed by moving the hands in front of the body, have a morphology which is totally different from spoken languages, which are produced by sequences of sounds. We improve the automatic alignment by applying a morphosyntactic and a semantic analysis and bridging the differences to find corresponding signs and phrases.
To handle the multimodality of sign languages in statistical machine translation, we present two approaches, focusing on the hands and on mouthings, i.e. silent lip movements pronouncing certain words in the sentence, as two modalities. In the first approach, we automatically adapt the granularity of the annotation by distinguishing signs with the same hand movements but different mouthings based on an automatic extraction of lip movements. In the second approach, we use the mouthing directly in the decoding process, using both the information signed by the hands and the mouthing as an input to the decoder.

By approaching the three issues of a suitable annotation, of data scarcity and of multimodality, we arrive at a translation system which can handle the multimodal sign language input and which is well beyond the performance of a standard translation system that only translates the manual component of a sign language utterance.


Da heutige Gebärdensprachkorpora sehr klein sind im Vergleich zu Korpora gesprochener Sprachen, führt dieser Mangel an Trainingsdaten oft zu einer schlechten automatischen Alignierung zwischen Gebärdensprachen und den entsprechenden Wörtern der gesprochenen Sprache. Dieses Problem wird noch durch die Bewegung der Hände ausgedrückt werden, eine ganz andere Morphologie haben als gesprochene Sprachen, die durch eine Folge von Lauten ausgedrückt werden. Wir verbessern die automatische Alignierung durch die Anwendung einer morphosyntaktischen sowie einer semantischen Analyse, um Entsprechungen zwischen Gebärdensprachen und gesprochenen Phrasen zu finden.

Zur Behandlung der Multimodalität von Gebärdensprachen in der automatischen Übersetzung präsentieren wir zwei Ansätze, wobei wir uns auf manuelle Gebärdensprachen und Mundbilder, also lautlos gesprochene Wörter, als zwei wichtige Modalitäten konzentrieren. Im ersten Ansatz passen wir automatisch die Granularität der Annotation an, indem wir Gebärdensprachvarianten mit gleichen manuellen Bewegungen, aber verschiedenen Mundbildern, unterscheiden. Hierzu extrahieren wir die Lippenbewegungen mit einem Active Appearance Model. Im zweiten Ansatz verwenden wir das Mundbild direkt im Übersetzungssystem, indem wir sowohl die Gebärdensprachen als auch das Mundbild als Eingabe des Übersetzungssystems verwenden.

Durch die Behandlung der drei Probleme einer geeigneten Annotation, des Trainingsdatenmangels und der Multimodalität der Gebärdensprache erhalten wir ein Gebärdensprachübersetzungssystem, welches als Eingabe multimodale Gebärdensprachen akzeptiert und welches eine wesentlich bessere Übersetzungsqualität hat als Systeme, die nur die manuelle Information der Gebärdensprachen verwenden.
Acknowledgements

First of all, I would like to thank Prof. Dr. Hermann Ney for giving me the opportunity to work and research at his chair in the field of statistical machine translation. I would also like to thank Prof. Dr. Ruben San-Segundo Hernández for accepting to be my second supervisor and for giving me a lot of valuable feedback on the draft version of this thesis.

My special thanks to Daniel Stein are two-fold: first for introducing me to and collaborating with me on the topic of sign language translation, and second for establishing the contact with the Fraunhofer IAIS where I work now. Next, I would like to thank Jens Förster, Oscar Koller and Yannik Gweth, my fellow colleagues in the EU FP7-project SignSpeak who worked on sign language recognition and with whom I collaborated on many publications and project reports. I am also thankful to Uwe Zelle and Horst Sieprath who taught me the basics of German Sign Language, stressing the importance of the usage of sign language space. In addition, several other members of the RWTH DESIRE group took part in the human evaluation of my translation system, which was very valuable for the research project SignSpeak.

Part of my duties at the chair was the administration of the chair finances, especially the external funds. I would like to thank Georg Heigold for teaching me the ins and outs of EU funding laws and Katja Bäcker and Gisela Gillmann as well as Andrea Starschich-Kierdorf and Stephanie Jansen for their continuous support, especially in times of financial audits. The financial administration surely was a lot of work, but I also learnt useful skills which I can still apply in my new position.

Among my colleagues, I would especially like to thank my office mates Gregor Leusch and Min-wei Feng as well as Malte Nuhn and Michal Kozielski for sharing many hours of work, exchanging tricks on UNIX tools and having lots of fun and laughter.

Last, but not least, I want to thank my head of department at Fraunhofer IAIS, Joachim Köhler, for encouraging me to finish this thesis.
## Contents

1 Introduction ................................................. 1  
1.1 Machine Translation .................................... 1  
1.2 Approaches to Machine Translation .................... 1  
1.3 Sign Languages .......................................... 2  
1.4 The SignSpeak Project .................................. 3  
1.5 About This Thesis ...................................... 4  

2 Scientific Goals ............................................ 5  
2.1 Goals .................................................. 5  

3 Preliminaries ................................................ 7  
3.1 Statistical Machine Translation ....................... 7  
3.1.1 Bayes Decision Rule ................................ 7  
3.1.2 Word-Based Machine Translation .................. 8  
3.1.3 Log-Linear Models .................................. 10  
3.1.4 Phrase-Based Machine Translation ................ 10  
3.1.5 Hierarchical Phrase-Based Machine Translation .... 13  
3.1.6 Shallow-1 Grammars ................................. 16  
3.1.7 Automatic Error Measures ........................... 17  
3.1.8 Optimisation ....................................... 18  
3.2 Sign Languages ......................................... 18  
3.2.1 Sub-lexical Components .............................. 19  
3.2.2 Signing Speed, Compactness and Incorporation .... 21  
3.2.3 Mouthing ........................................... 22  
3.2.4 Writing and Annotation Systems .................... 23  
3.3 Automatic Sign Language Recognition .................. 25  
3.4 Active Appearance Models ............................... 27  
3.5 Viseme Recognition ...................................... 30  
3.6 Summary ............................................... 34  

4 Related Work ................................................. 35  
4.1 European Research Projects .............................. 35  
4.2 Sign Language Corpora .................................. 35
## Contents

4.3 Sign Language Machine Translation ........................................... 37
4.4 Scarce Language Resources .......................................................... 39
4.5 Multimodality .............................................................................. 40

5 Annotation Schemata for Machine Translation ................................. 41
  5.1 Annotation Tools ......................................................................... 41
  5.2 The RWTH-PHOENIX-Weather Corpus ........................................ 42
     5.2.1 Motivation .......................................................................... 42
     5.2.2 Corpus Versions .................................................................. 44
     5.2.3 Annotation Scheme .............................................................. 45
     5.2.4 Corpus Setup ..................................................................... 46
  5.3 Efficient Annotation of Spoken Language Side .............................. 48
  5.4 Machine Translation System Setup .............................................. 50
  5.5 Experimental Results .................................................................. 51
     5.5.1 Effect of Recognition and Translation Errors ......................... 51
     5.5.2 Baseline Results RWTH-PHOENIX-Weather 2014 .................. 55
     5.5.3 Effect of Annotation on Translation ....................................... 56
  5.6 Summary and Contributions ....................................................... 57

6 Translation Using Scarce Resources ................................................. 59
  6.1 Improving Word Alignments Using Morphosyntactic Analysis ............ 59
  6.2 Using Synonym Data to Improve Alignments ................................. 62
  6.3 Additional Monolingual Data and Semi-Supervised Training .......... 63
  6.4 Summary and Contributions ....................................................... 67

7 Handling of Multimodality ................................................................. 69
  7.1 Clustering of Glosses According to Facial Features ......................... 69
     7.1.1 Sign Variants ..................................................................... 69
     7.1.2 Outline of the Algorithm ...................................................... 70
     7.1.3 Alignment .......................................................................... 71
     7.1.4 Active Appearance Models .................................................. 72
     7.1.5 Similarity Metric for Facial Expressions ................................. 72
     7.1.6 Clustering of Sign Variants .................................................... 73
     7.1.7 Evaluation .......................................................................... 73
     7.1.8 Summary ........................................................................... 74
  7.2 Incorporating Multimodality Into Decoding .................................... 75
     7.2.1 Mouthing Features in Machine Translation .............................. 76
     7.2.2 Setup of Multimodal System ............................................... 76
     7.2.3 Experiments ........................................................................ 77
     7.2.4 Summary ........................................................................... 79
  7.3 Corpus-NGT ................................................................................ 79
     7.3.1 Corpus Description ................................................................ 79
     7.3.2 Parallel Input Channel Approaches ....................................... 81
     7.3.3 Head Shakes ....................................................................... 82
     7.3.4 Summary ........................................................................... 84
  7.4 Summary and Contributions ....................................................... 84

8 Scientific Achievements and Conclusions ....................................... 87
  8.1 Scientific Achievements .............................................................. 87
Contents

8.2 Publications and Joint Work ............................................. 88
8.3 Conclusions ................................................................. 89

List of Figures ................................................................. 91
List of Tables ................................................................. 95
Bibliography ................................................................. 97
1.1 Machine Translation

The goal of machine translation is to convey the meaning of a text written in one natural language in another language. Such a translation is called fully automatic if the translation task is performed by a computer without any assistance of a human being.

1.2 Approaches to Machine Translation

Different approaches have been proposed to deal with this task of automatic translation. A general classification of the different approaches was presented by Vauquois [Vauquois 68] and summarised in the “Vauquois triangle”, a detailed version of which is shown in Figure 1.1.

In the diagram, the translation process is divided into two phases, an analysis phase and a generation phase. The approaches differ in the degree or level to which the source sentence is analysed. On one extreme is the interlingua approach in which the source sentence is transformed into an internal representation which captures the meaning of the sentence, from which the target sentence is then generated, following the grammatical rules of the target language. On the other extreme, in the direct approach the source sentence is directly transformed into the target sentence without resorting to an intermediate representation. The analysis can be done on different levels.

![Vauquois triangle](image)
such as the morphemic, syntactical or semantic level.

Different methodologies have been applied to the problem of machine translation, partly based on linguistic knowledge and hand-crafted models, partly on data-driven methods and statistical models.

In the rule-based approach, the source sentence is first analysed, and the syntactic structure is stored in a symbolic representation, e.g. a syntax tree. The computer then generates the translation by adjusting the syntactic structure and replacing words or phrases with their equivalents in the target language using a dictionary. The rules are usually created manually and thus can incorporate expert knowledge of a certain domain. This manual generation of rules becomes however increasingly complex for broader domains or even open domain systems, because the resulting rule set would have to handle an increasing number of exceptional cases or conflicts, which might be hard or impossible to resolve. On the other hand, rule-based approaches can be useful in comparison to data-driven approaches when no or only small amounts of training data is available. Before the emergence of contemporary statistical systems, a successful application of the rule-based approach was the commercial SYSTRAN system [Attnäs & Senellart+ 05].

When a new text is translated in the example-based approach, the system consults a bilingual text corpus and looks for similar sentences. It then translates the text in analogy to those sentences by combining the references found to a new whole. An example for this approach is the MATREX system developed by Dublin City University [Hassan & Ma+ 07].

The statistical approach uses probability theory to obtain the most probable translation of a sentence. One of the earliest publications were the word-based IBM-models [Brown & Della Pietra+ 93]. The statistical approach has several advantages: The system can be trained automatically, using a bilingual text corpus. A labour-intensive manual creation of linguistic rules is not necessary. When creating a translation system covering a new language pair, the system engineer does not even have to know any of the two languages. Moreover, the statistical approach can incorporate vague knowledge. While rule-based systems restrict sentences based on grammatical rules and thus prohibit the generation of sentences which contain grammatical errors, statistical systems use probabilities to describe which word sequences are likely or unlikely and can thus also deal with spontaneous speech which may contain grammatical errors.

In this work, we will apply the statistical approach to sign language machine translation. While we basically follow the direct approach, we do use a morphosyntactic analysis and a synonym database to mitigate the problem of scarce training resources. Sign language translation systems following the different approaches described above are presented in Section 4.3.

1.3 Sign Languages

Sign languages are the primal means of communication for most Deaf\(^1\) and many hard-of-hearing persons. They convey meaning by the movement and configuration of the hands, but also by other means of body language such as facial expressions, head and upper body movements, eye gaze and mouthings, i.e. the silent pronunciation of words or syllables. In sign languages, these means of expressions can be used sequentially, but also simultaneously. For example, a manual sign can be accompanied by a mouthing or a facial expression which can modify the meaning of the sign. In this work, we refer to this usage of different expressions as multimodality. One scientific challenge approached in this work is to feed this multimodal information as an input into a statistical translation system which usually only works on linear sequences of words.

\(^1\)A large proportion of the Deaf community actually considers itself to be part of a cultural minority rather than being impaired, differentiating the hearing condition “deaf” from the social group “Deaf” by a capital “D” (see [Rexroat 97]). We will adopt this notation in the following.
1.4 The SignSpeak Project

This thesis was written in the context of the research project “SignSpeak” and contributed to its overall success in developing a continuous sign language recognition and translation system.

The European Community’s Seventh Framework Programme STREP project “SignSpeak”² focused on the problem of automatic recognition and translation of continuous sign language. The aim of the project, which lasted from April 2009 to March 2012, was the development of a video-to-text system which converts a video of a signed utterance into the corresponding text of a spoken language. The technology could provide new e-services to the Deaf community, improving their communication with hearing persons [Dreuw & Forster+ 10].

The overall architecture of the system developed in the SignSpeak project can be found in Figure 1.2. It consists of two parts: first, a sign language recognition system is used to transform the video sequence into a symbolic representation of the signed utterance. Then, this representation is translated into a text in the spoken language. The representation chosen in the project consists of one or several streams of ID-glosses (see Section 3.2.4 for a description of ID-glosses). This representation is only used internally as an interface between the recognition and the translation system, that is, the user does not receive glosses as an output nor does he have to type in sequences of glosses as an input to the system.

It is clear that some information may be lost in the overall pipeline, partly because of error accumulation, partly because some information is not captured in the gloss representation and is thus not learned properly by the system. Consequently, the consortium enriched the gloss annotation of its corpora, the RWTH-PHOENIX-Weather and the Corpus-NGT, with additional information. In the case of the RWTH-PHOENIX-Weather, mouthings and mouth gestures, left hand signs, objects and locations of a movements are annotated, whereas in the Corpus-NGT head shakes and hand positions in the case of pointing gestures are annotated. Such information is important for a translation system to correctly translate negations, which are usually indicated by head shakes, and personal pronouns, which are indicated by pointing to specific locations in space which refer to previously mentioned persons or objects.

Figure 1.2: Architecture of the sign language recognition and translation system developed in the SignSpeak project

Note that in this work we contrast the term “sign language” with the term “spoken language” to refer to languages which are orally transmitted via sound patterns. We will however work with spoken languages in their written form, i.e. with texts. To be more concrete, our sign language to spoken language machine translation system will translate a video of the signed utterance to the text of the spoken language. We will not apply a speech synthesis system to create a spoken output.

²grant agreement no. 231424, see www.signspeak.eu
In the course of the project, we found that the Corpus-NGT was too challenging for the statistical methods of sign language recognition and translation due to its smaller corpus size and broader domain. We therefore concentrate our work in this thesis on the RWTH-PHOENIX-Weather Corpus. However, since the annotation of the Corpus-NGT consists of glosses for each individual hand and also additional head shake information, we present some experiments and an analysis of the corpus with respect to the multimodality of sign languages in Chapter 7.

1.5 About This Thesis

In this thesis, we enhance a statistical machine translation system to deal with the problem of translating from a sign language encoded in glosses to the text of the spoken language. We investigate the effect of the chosen sign language annotation scheme on the translation quality, deal with the problem of small training corpus sizes by incorporating morphosyntactic as well as semantic information and handle the issue of multimodality by including several modalities directly in the decoding process.

This work is structured as follows: Chapter 2 succinctly summarises the scientific goals pursued in this thesis. Chapter 3 introduces the basics of statistical machine translation, of sign languages and of active appearance models and viseme recognition which are used to detect mouthings. A brief survey of related work and the state-of-the-art in sign language translation is given in Chapter 4.

The main chapters of this work address the above mentioned goals: the effect of the annotation scheme used in the RWTH-PHOENIX-Weather 2014 corpus on the translation quality is analysed in Chapter 5, focusing on aspects such as mouthing and locations in the signing space. The issue of scarce corpus resources is dealt with in Chapter 6 using a morpho-syntactic analysis and synonym data. We address the challenges of multimodality in Chapter 7 by distinguishing sign variants using a clustering method, by merging multiple input streams and by combining multiple modalities in the translation process directly.

A conclusion and a summary of the scientific achievements of this work are given in Chapter 8.
In this chapter, we specify the scientific goals of this thesis and discuss the topics which will be covered in this work.

2.1 Goals

The following scientific goals were defined at the beginning and adjusted in the course of the work:

Related to sign language annotation:

- analyse sign language corpora and the used notation system with respect to their effect on statistical machine translation
- evaluate the effect of errors in the sign language recognition and translation system on the overall performance
- evaluate the impact of sign language recognition errors and sign language translation errors on human judgement
- develop an efficient way of annotating the spoken language side of a bilingual sign language corpus

Related to data scarcity:

- apply and adapt methods from machine translation to deal with the problem of scarce resources

Related to multimodality:

- deal with the problem of multimodality in sign languages: combine different information streams (hands, mouthing, head shake, etc.)
- discretise the multimodal structure of sign languages using clustering techniques
- handle the multimodal structure directly in the decoder
In this chapter, we will review the underlying concepts of statistical machine translation which form the basis of this work. In addition, we will discuss the difference between spoken and signed languages, which will become important when we deal with the translation of sign languages in later chapters. Since these topics have already been presented by other authors, we will only expound them briefly and refer the reader to the given literature for a more in-depth coverage. This chapter is organised as follows: The basics of statistical machine translation are described in Section 3.1. A brief introduction to sign languages is given in Section 3.2. The output of a sign language recognition system, which we introduce in Section 3.3, forms the input to our translation system. To automatically detect facial expressions and mouthings, we apply an active appearance model, which is outlined in Section 3.4. This model can be also be used to read the lips of the person signing, a technique called viseme recognition, which is covered in Section 3.5.

3.1 Statistical Machine Translation

In this work, we will mainly follow the approach of statistical machine translation which treats the machine translation task as a statistical decision problem. Simply put, the approach defines statistical models and then decides on the translation of a text or sentence based on these models. The parameters of the models are automatically learned from already translated texts, i.e. from a bilingual corpus. The statistical approach was firmly established with the development of the IBM models [Brown & Della Pietra 93]. Later, these word-based models were extended to phrase-based and hierarchical phrase-based models, which will be described in the following sections. To choose the best translation, usually the Bayes decision rule is applied.

3.1.1 Bayes Decision Rule

The task of machine translation is to translate a source language sentence $f^J = f_1 \ldots f_J$ of length $J$ into a target language sentence $e^I = e_1 \ldots e_I$ of length $I$. (The experiments in the groundbreaking paper by the IBM team were performed on the Canadian Hansards, i.e. the Canadian parliamentary speeches. After a parliamentary session, all speeches given either in French or English were translated into the respective other language, and thus a complete set of the proceedings in both French and English were produced. In the paper, the source language was therefore referred to as $f$ (“French”) and the target language as $e$ “English”. We will adhere to this convention in this work.)

According to the Bayes decision rule, the error at the sentence level is minimised if one chooses
the translation \( \hat{e}_1 \) with the highest posterior probability. \( \hat{e}_1 \) is therefore referred to as the maximum \textit{a posteriori} translation:

\[
f_1 \rightarrow \hat{e}_1(f_1) = \arg\max_{e_1} \{ Pr(e_1|f_1) \} \tag{3.1}
\]

In the source-channel approach, the posterior probability is reformulated using Bayes theorem:

\[
f_1 \rightarrow \hat{e}_1(f_1) = \arg\max_{e_1} \left\{ p(e_1|f_1) \right\} = \arg\max_{e_1} \left\{ \frac{p(e_1)p(f_1|e_1)}{p(f_1)} \right\} = \arg\max_{e_1} \left\{ p(e_1)p(f_1|e_1) \right\} \tag{3.2}
\]

Since in Equation (3.3) \( p(f_1) \) is independent of the maximizing variable \( e_1 \), it can be omitted. The resulting Equation (3.4) consists of two terms: \( p(e_1) \) is called the \textit{language model} and ideally measures the well-formedness of the sentence \( e_1 \), and \( p(f_1|e_1) \) is called the inverse \textit{translation model} and describes the probability that \( f_1 \) is the translation of \( e_1 \). Later in Section 3.1.3, the source-channel approach will be extended to a log-linear model combination which can include an arbitrary number of models.

### 3.1.2 Word-Based Machine Translation

Statistical machine translation gained momentum with the development of the IBM models in the early 1990s (early experiments can be found in [Brown & Cocke 1993, Brown & Della Pietra 1993]).

In the IBM models, the translation model \( Pr(f_1|e_1) \) is further decomposed. First, we introduce a hidden variable called the \textit{alignment} \( A \) which is a set of \textit{alignment points} modelling correspondences between words in the source language and the target language. For an example of such a word alignment see Figure 3.1. Note that in general, a word in one language can be aligned to none, one or multiple words in the other language and vice versa. Thus, an alignment in its most general form is neither a function of the source words nor the target words.

Without loss of generality, the translation model can be rewritten as:

\[
Pr(f_1|e_1) = \sum_A Pr(f_1, A|e_1) \tag{3.5}
\]

\[
Pr(f_1, A|e_1) = Pr(J|e_1) \cdot Pr(f_1, A|J, e_1) \tag{3.6}
\]

\[
Pr(J|e_1) \cdot Pr(A|e_1, J) \cdot Pr(f_1|e_1, J, A) \tag{3.7}
\]

The translation model is thus factored into a \textit{length model} \( Pr(J|e_1) \), an \textit{alignment model} \( Pr(A|e_1, J) \) and a \textit{lexical model} \( Pr(f_1|e_1, J, A) \).

The IBM team developed a sequence of models later referred to as the IBM models, which introduce additional model assumptions to make the model training feasible. In the following, we will give an outline of the IBM models, for a detailed description please refer to the original paper.

Since the IBM models do not have a closed form solution, their parameters are estimated using iterative methods and the expectation-maximisation (EM) algorithm. For these iterative methods, suitable starting values for the parameters have to be chosen. The IBM team therefore
3.1 Statistical Machine Translation

The European Commission will become overburdened if it pursues its current policy.

Figure 3.1: Alignment between a German sentence and its English translation. The black squares indicate correspondences between German and English words.

developed a sequence of models with increasing complexity and used the optimised parameters of each model as the starting value for the next model training.

Both IBM models 1 and 2 are zero-order models, i.e. the probability of each word in the sentence does not depend on the surrounding words. In contrast to the general alignment $A$, the above models assume that each source word $f_j$ is aligned to at most one word $e_i$ in the other language. Thus, for each source position $j$ an alignment point $a_j = i$ is stored. Here, $a_j = 0$ indicates that the source word is not aligned to any target word. The whole alignment $A$ is then the combination of all alignment points: $A := a_1, \ldots, a_j, \ldots, a_J$. This leads to a great reduction in the number of possible alignments from $2^J$ to $(I + 1)^J$. While in IBM model 1 the alignment is uniformly distributed, IBM model 2 assumes an alignment model which depends on the source position.

IBM model 3 also allows one word to produce several words in the other language. The fertility $\phi$ is the number of words produced by one word.

IBM models 4 and 5 introduce more detailed dependencies for the alignment model, particularly reordering probabilities based on relative positions and word classes. Since IBM model 4 is deficient, i.e. not normalised, IBM model 5 was introduced to correct the deficiency.

Originally, the IBM models were used for a word-based stand-alone translation system. State-of-the-art systems are usually phrase-based or hierarchical phrase-based systems, since these systems perform much better than word-based systems. However, the IBM models can also be used to obtain word alignments from a bilingual corpus, which serves as a starting point for the more advanced methods. Since the alignments of these models are a function of one language, they are not symmetric. Consequently, training the alignment in one direction leads to different results than training the alignment in the other direction. In practice, alignments for both directions are trained, and the resulting two alignments are merged using certain heuristics [Och & Ney 03].

In state-of-the-art translation systems, the alignments are usually trained using the open-source program GIZA++ [Och & Ney 03]. In addition to the IBM models described in the original paper, a hidden Markov model (HMM), which is a first-order model, is introduced between IBM-2 and IBM-3 [Vogel & Ney + 96].
3 Preliminaries

3.1.3 Log-Linear Models

In the source-channel approach (Equation (3.4)), the posterior translation probability is reformulated into a language model and a translation model. In order to add an arbitrary number of additional knowledge sources into a translation system, this equation is generalised to a log-linear model:

\[
p(e_I^1|f_J^1) = \frac{\exp \left( \sum_{m=1}^{M} \lambda_m h_m(f_J^1, e_I^1) \right)}{\sum_{\tilde{e}_I^1} \exp \left( \sum_{m=1}^{M} \lambda_m h_m(f_J^1, \tilde{e}_I^1) \right)}
\] (3.8)

The functions \(h_m(f_J^1, e_I^1)\) are called feature functions and can model arbitrary properties of a translation \(e_I^1\). Each feature function has a scaling factor \(\lambda_m\). These factors are optimised on a held-out portion of the data, the development set. Note that the source-channel approach is a special case of the log-linear model with \(M = 2\), \(h_1(f_J^1, e_I^1) = \log(p(e_I^1))\), \(h_2(f_J^1, e_I^1) = \log(p(f_J^1|e_I^1))\), \(\lambda_1 = \lambda_2 = 1\).

When applying the Bayes decision rule (Equation (3.1)) to the log-linear model (Equation (3.8)), we obtain:

\[
f_J^1 \rightarrow \hat{e}_I^1(f_J^1) = \arg\max_{e_I^1} \{ p(e_I^1|f_J^1) \}
\] (3.9)

\[
= \arg\max_{e_I^1} \left\{ \frac{\exp \left( \sum_{m=1}^{M} \lambda_m h_m(f_J^1, e_I^1) \right)}{\sum_{\tilde{e}_I^1} \exp \left( \sum_{m=1}^{M} \lambda_m h_m(f_J^1, \tilde{e}_I^1) \right)} \right\}
\] (3.10)

\[
= \arg\max_{e_I^1} \left\{ \exp \left( \sum_{m=1}^{M} \lambda_m h_m(f_J^1, e_I^1) \right) \right\}
\] (3.11)

\[
= \arg\max_{e_I^1} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f_J^1, e_I^1) \right\}
\] (3.12)

The denominator in Equation (3.10) is again independent of the maximizing variable \(e_I^1\) and can be omitted. Since the exponential function is monotonic, it can be omitted inside the \(\arg\max\)-function as well.

Figure 3.2 shows a system architecture using a log-linear model. A translation system which uses the log-linear model can be easily extended by introducing new feature functions. The following sections describe two translation paradigms, the phrase-based and the hierarchical phrase-based systems, which each introduce several additional feature functions into the log-linear framework.

3.1.4 Phrase-Based Machine Translation

One disadvantage of the word-based IBM translation models is that the lexical translation probability \(p(f_j|e_a)\) of a word \(f_j\) only depends on the word \(e_a\) it is aligned to but does not take the surrounding context of the sentence into account. To overcome this issue, phrase-based models [Och & Tillmann + 99, Zens & Och + 02, Koehn & Och + 03] extend the basic unit of translation from words to contiguous phrases. Note that in this context the word “phrase” is not used in a linguistic sense but simply refers to a sequence of contiguous, i.e. adjacent words in a sentence.

Bilingual phrases, i.e. phrases in the source sentence and their corresponding translation in the target sentence, are extracted from a bilingual training corpus. First, the training corpus has
3.1 Statistical Machine Translation

Figure 3.2: System architecture using a log-linear model combination. The system is easily extensible by adding additional feature functions.
Die Europäische Kommission wird mit Arbeit überlastet, wenn sie diezeitige Politik weiterverfolgt.

The European Commission will become overburdened if it pursues its current policy.

Figure 3.3: Alignment between a German sentence and its English translation. The blue dashed rectangles indicate some valid phrase pairs. Note that valid phrases can overlap or contain other valid phrases.

to be word aligned. This can be done using the IBM-models with the publicly available GIZA++ toolkit (see Section 3.1.2). Phrase pairs \( \langle \tilde{f}, \tilde{e} \rangle \) of contiguous phrases \( \tilde{f} \) and \( \tilde{e} \) in the source and target sentence are extracted if the words of \( \tilde{f} \) are only aligned to words of \( \tilde{e} \) and vice versa.

The set of phrase pairs of a given sentence pair \( (f^J_1, e^I_1) \) with the word-alignment \( A \) is defined as:

\[
\mathcal{BP}(f^J_1, e^I_1, A) = \left\{ \langle f_{j_1}^{j_2}, e_{i_1}^{i_2} \rangle \mid 1 \leq j_1 \leq j_2 \leq J, 1 \leq i_1 \leq i_2 \leq I, \right. \\
\left. \forall (j, i) \in A : j_1 \leq j \leq j_2 \Leftrightarrow i_1 \leq i \leq i_2 \right. \\
\left. \land \exists (j, i) \in A : j_1 \leq j \leq j_2 \land i_1 \leq i \leq i_2 \right\}
\]

(3.13)

Taking the example sentence of Figure 3.1, Figure 3.3 shows some valid bilingual phrase pairs according to the above definition. For example, single word phrases such as \( \langle \text{Die}, \text{The} \rangle \), \( \langle \text{Europäische}, \text{European} \rangle \), but also longer phrases such as \( \langle \text{Die Europäische Kommission}, \text{The European Commission} \rangle \) are valid phrase pairs. In the example, for the target phrase “if it pursues” no valid bilingual phrase can be extracted, since the corresponding source phrase “wenn sie weiter verfolgt” is not contiguous but would contain a gap. We will later see that such phrase pairs are used in the hierarchical phrase-based approach.

In the training phase, all valid phrases are extracted for all training sentences, and their translation probabilities are estimated using relative frequencies. In the translation of a test set, i.e. a new text unseen by the system, a sentence is first segmented into phrases, which are then translated into the target language using the bilingual phrases. Moreover, the phrases can be reordered to allow for a different word order in the target language.

The phrase-based system evaluates each translation using several feature functions in a log-linear model (Equation (3.12)). These features include:

- **Phrase translation probabilities in both directions**

  Both phrase translation probabilities from source to target and from target to source are
3.1 Statistical Machine Translation

- **Word translation probabilities in both directions**
  Another method to correct unstable estimates for rare phrases is to use lexical weighting. Lexical weighting can be defined as in the IBM-1 model and can be regarded as a smoothing method.

- **Phrase reordering probabilities**
  Languages differ in the order in which words appear in a sentence. Since the computer does not know the correct word order, it tries different permutations of the source phrases and chooses the translation with the best log-linear model score. Reordering models assign scores to these different permutations. A variety of reordering models exist, ranging from simple distance models to lexicalised models which learn that certain words or phrases are reordered more often than others.

- **One or several language models**
  Language models estimate the probability of the target sentence. They are used to judge the well-formedness of the translation. Usually, an \( n \)-gram approach is taken, where the probability of the sentence is split into the product of each individual word given the history of the preceding \( n - 1 \) words. Smoothing techniques are applied to estimate the probability of unseen \( n \)-grams. In this thesis, we use Kneser-Ney smoothing [Kneser & Ney 95].

- **Word and phrase penalties**
  Since the language model as well as the word and phrase translation probabilities are defined as products over individual words and phrases, longer translations usually have a smaller probability than shorter translations, and a translation system would therefore prefer shorter translations. To counteract this problem, word and phrase penalties are introduced which score a translation based on the number of words and phrases it contains.

3.1.5 Hierarchical Phrase-Based Machine Translation

One restriction of the phrase-based paradigm presented in the previous section is that phrases have to consist of contiguous words, i.e. words adjacent to each other without any gaps. However, most languages contain recursive constructions where a phrase can contain another phrase. One example for such a recursive structure are German separable verbs. In Figure 3.4, the finite form of the separable verb “ankommen” (“to arrive”) is separated into “kommt” and “an” and frames the prepositional phrase “am Samstag” (“on Saturday”). The phrase-based paradigm can only learn the whole phrase pair (“kommt am Samstag an”, “arrives on Saturday”), because phrases have to be contiguous. In the hierarchical approach, the more general phrase (“kommt \( X \) an”, “arrives \( Y \)”) can be learned, where \( X \) and \( Y \) are placeholders for other phrases in the source and target language which are translations of each other. In the hierarchical approach, phrases do not have to be contiguous, and long-range dependencies can be captured better than in the phrase-based approach. Moreover, hierarchical phrases can also incorporate reorderings, as can be seen in Figure 3.5. In German relative clauses, the verb is placed at the end of the clause, while English uses the regular subject-verb-object structure (Figure 3.5a). The hierarchical phrase in Figure 3.5b can perform the necessary reordering. It constitutes a form of lexicalised reordering, as it recognises the relative clause based on the words “, which” and “, die”.

The hierarchical phrase-based approach is formally modelled using synchronous context free grammars (SCFGs) [Lewis II & Stearns 68], which are a generalisation of context free grammars.
(a) Example sentence containing the separable verb "ankommen" (to arrive).

(b) A hierarchical phrase for the separable verb "ankommen".

Figure 3.4: Separable verbs in German are one example of brace structures which cannot be learned in the phrase-based approach.

(a) Example sentence showing the different word order of relative clauses in German and English.

(b) A hierarchical phrase extracted from the sentence in 3.5a. The phrase moves the German verb phrase $X^\sim 1$, which is in the last position, in front of the adverbial phrase "in $X^\sim 0$".

Figure 3.5: Hierarchical phrases can capture different sentence structures. In German relative clauses, the verb is in the last position, while English subordinate clauses follow the regular subject-verb-object structure.
3.1 Statistical Machine Translation

[Chomsky 56] applied to two languages at the same time. This grammar was originally developed to model compiler translations but can be applied to the translation of natural languages as well. In context-free grammars, non-terminal symbols can be replaced by other symbols by applying so-called production rules, while terminal symbols cannot be further replaced by other symbols. Thus, in the grammar sub-phrases can be inserted into larger phrases by replacing a non-terminal symbol with the sub-phrase.

The non-terminal symbols in the SCFG capture phrase-level reorderings, while the order of the terminal symbols on the source and target side can model word-level reorderings.

A SCFG consists of the tuple $(\mathcal{F}, \mathcal{E}, \{S_X, X\}, \{S_Y, Y\}, \mathcal{H}, (S_X, S_Y))$ with

- the set of source words $\mathcal{F}$ (the alphabet of terminal symbols on the source side)
- the set of target words $\mathcal{E}$ (the alphabet of terminal symbols on the target side)
- the set of non-terminal symbols $X$ on the source side, including the start symbol $S_X$
- the set of non-terminal symbols $Y$ on the target side, including the start symbol $S_Y$
- the set of production rules $\mathcal{H}$, corresponding to hierarchical and lexical phrases
- the start symbol pair $(S_X, S_Y)$

The set of production rules basically consist of two types:

- **Lexical rules**
  A phrase pair $(f, e)$ in the phrase-based approach corresponds to a lexical rule $X \rightarrow (f, e)$.

- **Hierarchical rules**
  A hierarchical rule is constructed by removing a lexical phrase from a larger phrase: if the above phrase pair is removed from the rule $X \rightarrow (\alpha_1 f \alpha_2, \beta_1 e \beta_2)$, where $\alpha, \beta \in \mathcal{F} \cup \mathcal{X}, \beta \in \mathcal{E} \cup \mathcal{Y}$, the resulting hierarchical rule is $X \rightarrow (\alpha_1 X_1 \alpha_2, \beta_1 X_1 \beta_2)$. Usually, the number of non-terminal symbols on the right-hand side of a rule is restricted to two. It can be shown that such a grammar is equivalent to an inversion transduction grammar [Wu 97].

In practice, the set of non-terminal symbols is set to $\{S, X\}$ for both languages. In addition to the above types of rules, two additional rules are added. The **start rule** $(S, S) \rightarrow (X, X)$ converts the start symbol to the standard non-terminal $X$. The **glue rule** $S \rightarrow (SX, SX)$ allows for the monotonic concatenation of phrases as in the phrase-based approach. **Paste rules** of the form $(X, X) \rightarrow (X^\langle 0 \rangle \alpha, X^\langle 0 \rangle \beta)$ or $(X, X) \rightarrow (\alpha X^\langle 0 \rangle, \beta X^\langle 0 \rangle)$ add phrases to the left or the right of the hierarchical tree, and thus no reordering occurs.

In a weighted SCFG, each rule is assigned a weight, and the weight of a derivation is the product of the weight of the rules which were used for the derivation. As in the phrase-based approach, the derivation weight is combined with other scores in a log-linear model.

The features of the hierarchical model include all the features described in the phrase-based approach except for the reordering model. Additionally, the following features are used:

- **Hierarchical rule**
  binary feature indicating that the rule is a hierarchical rule. The feature weights the use of hierarchical rules vs. lexical rules.

- **Glue rule**
  binary feature for the glue rule $S \rightarrow (SX, SX)$. 

15
Derivations for $X \sim 0$

Figure 3.6: Illustration of the cube pruning algorithm (taken from [Vilar & Ney 12]), for a hyperedge with two predecessors. Each axis corresponds to each element that contributes to the total score: the derivations associated with each of the non-terminals and the possible rules (translations) in the hyperedge. The 3 lightly shaded cubes correspond to the 3-best derivations along the hyperedge and have been generated in order of increasing costs. The dark shaded cubes correspond to the active candidates for the next-best derivation.

- **Paste rule**
  
  binary feature for rules of the form $(X, X) \rightarrow (X \sim 0 \alpha, X \sim 0 \beta)$ or $(X, X) \rightarrow (\alpha X \sim 0, \beta X \sim 0)$

For decoding, a modified CYK+ [Chappelier & Rajman 98] parser is used, which is an extension of the standard CYK parser for context free grammars. The CYK+ parser relaxes the requirement of the grammar to be in Chomsky normal form. A problem which is not straightforward is the integration of the language model feature into the parser and the efficient pruning of hypotheses with unpromising scores. This problem is solved with the cube pruning algorithm [Chiang 07]. An illustration of the algorithm can be found in Figure 3.6. For more information please refer to the above paper and to [Vilar & Ney 12].

The hierarchical phrase-based approach is more general than the phrase-based approach, learning general hierarchical rules. To learn these hierarchical rules reliably, a large amount of training data is necessary. Since sign language corpora are usually much smaller than bilingual corpora for spoken languages, one research question is whether hierarchical systems can be reliably trained.

### 3.1.6 Shallow-1 Grammars

As described in the previous section, hierarchical phrase-based systems are more general than phrase-based systems by introducing hierarchical rules which contain non-terminal symbols on the right hand side and which thus allow recursive structures, although the search space of the hierarchical system is not a superset of the phrase-based system because of the different reordering approach of the phrase-based system.
3.1 Statistical Machine Translation

It is possible to restrict the depth of the recursion in a hierarchical-system by modifying the grammar, creating shallow-\(n\) grammars with a maximum depth of \(n\) [de Gispert & Iglesias+ 10]. The resulting system decodes more quickly and prevents over-generation.

In a shallow-1 grammar, the non-terminal \(X\) of the hierarchical grammar is replaced by two distinct symbols \(X_H\) and \(X_L\), representing hierarchical and lexical phrases, respectively. Thus, on the left-hand side of all hierarchical rules, the non-terminal symbol is set to \(X_H\), on all lexical rules to \(H_L\). On all right hand sides, the non-terminal \(X\) is replaced by \(X_L\), and thus hierarchical rules cannot be inserted into other hierarchical rules. The start rule is substituted with 
\[
S \rightarrow \langle X_H, X_H \rangle
\]
and 
\[
S \rightarrow \langle X_L, X_L \rangle
\]
and the glue rule is replaced with 
\[
S \rightarrow \langle S^\sim 0 X_L^{-1}, S^\sim 0 X_H^{-1} \rangle
\]
and 
\[
S \rightarrow \langle S^\sim 0 X_H^{-1}, S^\sim 0 X_L^{-1} \rangle
\].

The phrase-based, the hierarchical phrase-based and the shallow-1 approach mainly differ in their degree of generalisation. While the phrase-based system only allows for a reordering of contiguous phrases, the hierarchical system also uses non-contiguous phrases. The shallow-1 approach restricts the hierarchical approach by allowing only contiguous phrases in the gaps of non-contiguous phrases and thus can be considered a middle ground between the phrase-based and the hierarchical phrase-based approach. The question which system performs best for the task of sign language translation mainly depends on the corpus size, because for the reliable estimation of hierarchical phrase probabilities, a larger corpus is needed than for the estimation of contiguous phrases. In our baseline experiments in Section 5.5.2, we compare the three approaches on the RWTH-PHOENIX-Weather 2014 corpus and obtained the best translation performance using the shallow-1 approach.

3.1.7 Automatic Error Measures

There are two ways to evaluate the quality of a translation: a human evaluation and an automatic error measure. In the human evaluation, a person proficient in both languages judges the quality of the translation. Note that often a sentence can have more than one possible translation, and sometimes translations differ only in style or register. Thus, sometimes the judgement of the translation quality also differs between different persons or even between judgements of the same person done at different times.

Automatic quality metrics are usually based on one or several reference translations, comparing the system output to these references. In contrast to the subjective human judgement, they are objective in the sense that a given reference translation and hypothesis will always lead to the same automatic quality measure, while a human judgement may vary. Another advantage of automatic metrics is that they can be computed fast, while human evaluation is a time-consuming and laborious effort.

The following metrics are commonly used in statistical machine translation and have been applied in this work:

**WER** The “Word Error Rate” is defined as the Levenshtein distance [Levenshtein 66] between the hypothesis and the reference translation, which is defined as the minimum number of insertions, deletions and substitutions of words to transform the former into the latter.

**TER** The “Translation Edit Rate” [Snover & Dorr+ 06] is an extension of the word error rate which allows shifting whole phrases in addition to the above transformations. Since e.g. temporal clauses can often be placed in different positions in the sentence without altering their meaning, such an operation is quite natural, and consequently TER has replaced WER as a common evaluation metric in the field of machine translation.

**BLEU** BLEU stands for “BiLingual Evaluation Understudy” [Papineni & Roukos+ 02] and calculates the \(n\)-gram precision between the hypothesis and the reference. Since precision metrics
tend to overrate short hypotheses, a brevity penalty was introduced. Note that in contrast to the above error metrics, higher BLEU values indicate a better translation.

In our experiments, we will present results both in BLEU and TER. Parameter optimisation is usually done on BLEU unless specified otherwise. Besides the automatic metrics, we also performed a human evaluation in Section 5.5.1.

### 3.1.8 Optimisation

In the log-linear model in Equation (3.8), we introduced free parameters $\lambda_m$ to give different weights to the feature functions. The weights $\lambda_m$ have to be optimised on a separated part of the training data called the development set. Ideally, the development set should have the same distribution of words and phrases as the new text we want to translate, which is called the test set.

The parameters are optimised using Franz Och’s method which is usually referred to as minimum error rate training [Och 03]. The method optimises the parameters on an $n$-best list of hypothesis translations of the development set. Starting with a set of parameters to generate the $n$-best list, one optimisation step obtains a new set of parameters, which can again be used to generate an $n$-best list, leading to an iterative process.

Och’s method optimises one scaling factor at a time in random order. Since the cost function used to select the best hypothesis is linear in the weights $\lambda_k$ (see Equation (3.12)), the cost function becomes a line with regard to a single $\lambda_k$ and one hypothesis $e_1^f$:

$$c(\lambda_k) = \lambda_k h_k(e_1^f, f_1^f) + \sum_{m=1, m \neq k}^{M} \lambda_m h_m(e_1^f, f_1^f)$$  (3.14)

For the $n$-best list of hypotheses, this leads to a set of $n$ linear equations. Since the hypothesis with the lowest cost is chosen, one has to calculate the lower envelope of the $n$ lines, which can be done using the sweep line algorithm [Bentley & Ottmann 79]. The error measure has to be computed only at the intersection points of the lines, as the intersections form the decision boundaries. The parameter leading to the best error measure is then chosen, and the algorithm continues to optimise the next parameter.

Since the calculation of the error metric has to be performed hundreds or thousands of times in the course of the iterative optimisation process, using a human evaluation of the generated hypothesis is infeasible. We therefore need an automatic metric such as BLEU or TER.

As sign language corpora are considerably smaller than spoken language corpora, withholding a dedicated development set already strips away a considerable portion of the training data. In [Stein & Schmidt + 12, Section 6.1], we therefore opted for using a technique similar to n-fold cross-validation for optimisation. We split our training set into a number of equally sized sets (in this work 6 sets). Then, we train six different translation systems, each time withholding one of the sets. In each optimisation iteration, we concatenate the $n$-best lists of all individual systems for a complete training set translation. See Figure 3.7 for a graphical representation of this method. For contrastive results with and without this technique, see Section 5.5.2.

### 3.2 Sign Languages

Sign languages are the primal means of communication for most Deaf and many hard-of-hearing persons. They convey meaning visually by manual communication and body language, in contrast to spoken languages which are acoustically conveyed via sound patterns. Sign languages differ
3.2 Sign Languages

Figure 3.7: Graphical representation of the applied optimisation methods. The standard method withholds a part of the training data as development set. The improved method combines several systems, each withholding a different development set, similar to the method of n-fold cross-validation.

from country to country, and often several regional dialects exist within one country. Most sign languages have evolved naturally due to the need of Deaf persons to communicate and particularly exploit the visual modality by depicting concrete objects or actions. Thus, they considerably differ in grammar and vocabulary from the spoken language of the same country or region. One feature which is unique to sign languages is the simultaneous conveyance of meaning via multiple information channels: the two hands, facial expression and mouthing, body posture, head movement, eye gaze, etc. While spoken languages are conveyed via a linear sequence of phonemes, sign languages can use all of these components simultaneously. By modifying one or several of the components, the meaning of a sign can be partly altered or adapted, leading to a highly inflected language.

In the following sections, we will briefly introduce some of the core features of sign languages which are relevant to the task of machine translation. For a more thorough introduction, please refer to [Neidle & Kegl + 00] for American Sign Language (ASL) and [Braem 95] for German Sign Language (“Deutsche Gebärdensprache”, DGS).

3.2.1 Sub-lexical Components

Sign languages express meaning by visual patterns. More precisely, the signer uses several articulators to express meaning simultaneously, see Table 3.1.

While in spoken languages some of these means are also used for non-verbal communication, e.g. expressing emotions by facial expressions or shaking one’s head to indicate disagreement, in sign languages these means are part of the grammatical system. For example, some statements in a sign language can be negated solely by shaking one’s head while signing.

In spoken languages, phonemes are the smallest contrastive unit of sounds in a language which bring about a change in meaning. Phonemes can be recognised by comparing two words which only differ in one sound, e.g. the words “butter” (a dairy product) and “putter” (a golf club). This minimal pair indicates that /p/ and /b/ are two different phonemes in English. Studies in sign languages (see [Braem 95], pp. 18-26 and [Neidle & Kegl + 00], p.28) show that hand shapes, orientations, movements, and positions are the equivalent of phonemes. For example, Figure 3.9 shows an example of two signs differing only in the used hand shape. Thus, the “1-handshape” and
Table 3.1: Sign language articulators which can be used simultaneously

<table>
<thead>
<tr>
<th>manual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>hand shape</td>
<td>configuration of the fingers, e.g. pointing finger, flat hand, o-shape, etc. See also Figure 3.9.</td>
</tr>
<tr>
<td>hand orientation</td>
<td>palm facing upwards, downwards, to the front, etc.</td>
</tr>
<tr>
<td>hand position</td>
<td>touching certain body parts such as the head, the chest, or holding the hand in front of the body</td>
</tr>
<tr>
<td>movement of hands and</td>
<td>Many signs include motion, e.g. to sign the word “house”, the shape of the roof and walls are traced with flat hands.</td>
</tr>
<tr>
<td>arms</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>non-manual</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>facial expressions</td>
<td>facial expressions can express emotions or degree/amount, e.g. in DGS, the notion of “large” is implied by puffing one’s cheeks and opening one’s eyes widely, see Figure 3.8.</td>
</tr>
<tr>
<td>eye gaze</td>
<td>The signer usually looks at his conversation partner but can also look to the signing space in front of him to indicate importance or focus.</td>
</tr>
<tr>
<td>head movement</td>
<td>Sentences can be negated simply by shaking one’s head. Nodding can affirm a statement.</td>
</tr>
<tr>
<td>upper body posture</td>
<td>The upper body can be used e.g. to repeat a conversation of two people, indicating one person by slightly turning to the left, the other by turning to the right.</td>
</tr>
<tr>
<td>mouthing,</td>
<td>Some words can be accompanied by mouthing, i.e. the silent pronunciation of a word, see Section 3.2.3. Mouth gestures are often used to indicate motion and action, e.g. a blowing gesture to indicate wind.</td>
</tr>
<tr>
<td>mouth gestures</td>
<td></td>
</tr>
</tbody>
</table>
3.2 Sign Languages

(a) Puffed cheeks imply the notion “large, heavy”
(b) Tight round lips imply the meaning “small, tiny, thin”.

Figure 3.8: Facial expressions can express degrees/amounts. The examples are in DGS, figures taken with permission from www.visuelles-denken.de

the “F-handshape” can be considered a minimal pair, and the handshape component a “phoneme”. William Stokoe, who was one of the earliest scholars to introduce an annotation system for sign languages (see Section 3.2.4), used the term “chereme”, however contemporary linguistics usually also applies the term “phoneme” to sign languages.

3.2.2 Signing Speed, Compactness and Incorporation

[Bellugi & Fischer 72] showed that on average a sign in ASL takes longer to produce than a spoken word in English, but that the production of a proposition takes roughly the same amount of time in both languages. The higher rate of speech production can be explained by the smaller movements of the speech apparatus (glottis, tongue, mouth, lips) when compared to the complex and larger movements of the hands and arms in the signing space in front of the body when producing a sign. We can confirm this finding, as the corpus statistics of our parallel sign language corpus (see Section 5.2.4) show that the number of signs used to form a proposition is smaller than the number of spoken words to express the same proposition: in the corpus of 8,834 sentences, the total number of raw glosses without additional information is roughly 74k, the number of spoken words excluding punctuation marks is 122k.

In the above paper, Susan Fisher expounds the mechanisms by which sign languages compensate for the fact that signs take longer to produce than spoken words and enumerates three characteristics.

The first strategy is to condense the message by removing redundancy and eliminating function words which are frequently used in spoken languages. E.g. the English sentence “It is against the law to drive on the left side of the road.” (example from [Bellugi & Fischer 72, p.14]) is condensed to the ASL signs “ILLEGAL DRIVE LEFT-SIDE”. This implies that when translating from a sign language to a spoken language, these function words have to be added by the system.

Secondly, sign languages can incorporate various aspects such as direction of verbs, number of persons, manner, size and shape of the subject or object into a sign by slightly modifying parameters of the sign. E.g., the sign “TO-GIVE” incorporates the subject and object in its direction. To sign “I give to you”, the signer would move the giving hand from himself to the other person, while “You give to me” is signed by moving the hand from the spectator towards himself. Similarly, the manner in which an action happens can be incorporated into the action by modifying it. For example, to indicate that a car can hardly make it up a hill, the sign “TO-DRIVE” would be signed very slowly with a grim face and tense muscles to indicate that it is
difficult for the car to progress. Incorporation poses a difficulty for the notation system: how can these modifications of a sign be annotated in sufficient detail?

Thirdly, sign languages can express various aspects such as questions, the person speaking during direct speech, discourse management, etc. with facial expressions or body movements. E.g., yes-no questions in DGS are indicated by raising one’s eyebrows at the end of the proposition, affirmative sentences are indicated with a nod, while a negative sentence is indicated with a head shake.

Spoken languages use anaphora such as pronouns (e.g. he, she, they) and relative clauses (e.g. which, who) to concisely refer to already mentioned persons and objects. Sign languages usually place these objects in an imaginary space, the token space, and later refer to them by pointing to the same position in space. When one wants to retell a conversation of two persons, one can put one person to the left and one person to the right in the token space and refer to these two persons by pointing to the left and right position, respectively.

One particular feature of incorporation in sign languages are classifiers. They are used to identify the position of objects or persons, trace contours to depict even complex forms or draw paths through space to indicate movements. These predicates are constructed using one of a closed set of handshapes which characterises the entity (thus the handshape “classifies” the entity) and performing possibly complex movements with this handshape. For a detailed discussion of classifiers and their complexity see [Huenerfauth 04b]. In the domain of weather forecasting, which we will be mostly concerned with, classifiers are used to indicate the position of a region on the map or to depict the movements of rain clouds or high pressure areas.

3.2.3 Mouthing

One non-manual aspect of sign languages is mouthing. Some countries such as Germany have a strong tradition of oralist education, i.e. the teachers tried to educate Deaf children through oral language. Deaf students had to learn lip reading and tried to mimic speech in spite of not hearing the sounds they produce. The main idea behind oralist education is that Deaf people are better integrated into the hearing community by good skills in lip reading and production of speech,

Figure 3.9: An example of two signs in German Sign Language differing only in their handshape. Left meaning: “to say” (“1-handshape”), right meaning: “to ask” (“F-handshape”) (taken from [Braem 95],p.19)
3.2 Sign Languages

Figure 3.10: The ASL sign “line” as an illustration and in Stokoe notation. (Dictionary entry from [Byrom 08], p. 60) The notation $\bar{B}_aG_x^\perp$ indicates: the left hand has a flat hand shape ($B$) facing up ($a$) and is under ($\bar{\;}$) the right hand, which forms a pointing finger ($G$) touching ($x$) the left palm and moving away ($\perp$) from the signer.

which usually sounds like whispering. The disadvantage is that only about 30% of the spoken language can be recognised by lip reading, because several phonemes look the same (see more information about phonemes and visemes in Section 3.5). Thus, oralist education is very difficult and frustrating for Deaf children, and children educated in sign language as their first language and the written form of a spoken language as a second language (bilingual education) can learn the curriculum faster. See [Hermans & Knoors+ 08] for a discussion on the topic.

In countries with a strong oralist tradition, sign languages tend to frequently use mouthings, that is, mouth patterns derived from the spoken language. In these languages, mouthings are also used to derive new signs by using the manual components of a similar sign and changing only the mouthing. Note that not all signs are accompanied by mouthings but only a few specific words. In German Sign Language, the derivation of new nouns by a mouthing is particularly common, while verbs are hardly combined with mouthings but rather with mouth gestures. For example, actions are often accompanied by blowing lips to indicate motion. In this thesis, we use a mouthing recognition system to improve the standard recognition and translation system which is mainly based on the manual components.

3.2.4 Writing and Annotation Systems

Sign languages lack an official writing system. Annotation systems are mainly used in the academic context of linguistic studies. These systems vary greatly in their level of detail, depending on the purpose of the annotation scheme. One of the earliest annotation systems was presented in [Stokoe 60], which decomposes a sign into a set of subunits similar to phonemes in the spoken language. A more detailed system is the HamNoSys (Hamburg Notation System, [Prillwitz 85]), which is widely used in academia nowadays. An example for the Stokoe notation and HamNoSys can be found in Figures 3.10 and 3.11, respectively.

A different approach was taken by Valerie Sutton, who developed a writing system for sign languages called SignWriting [Sutton & Writing 00]. Sutton, who is a dancer, has adapted the system from DanceWriting, a notation system for dance. SignWriting is highly visually iconic and thus easy to understand for a person who knows the sign language to be annotated. It is

---

1Entry taken from the University of Hamburg online sign language dictionary: http://www.sign-lang.uni-hamburg.de/projekte/slex
Figure 3.11: The sign “to say” (German: “sagen”, see Figure 3.9 left) in HamNoSys notation. Symbols from left to right: The handshape depicts a hand with pointing index finger (left), which can alternatively be bent (right). The orientation indicates that the hand is pointing upwards and the palm is facing diagonally to the right of the signer. The location is at the chin (left), and in contact (right) with the chin. The action indicates a movement forward and downward (1. symbol) in a circular movement (2. symbol) which is due to an elbow rotation. After (3. symbol) the downward movement, the index finger is pointing diagonally to the left (4. symbol).

Figure 3.12: The sign “question” (German: “Frage”) in SignWriting notation. The four circles indicate a temporal sequence of the mouthing of the word “f-r-a-g-e”. The handshape is iconic and similar to the drawing in Figure 3.12. The arrow indicates a movement horizontally away from the mouth, the horizontal direction is encoded in the colour of the arrowhead and its stroke. The initial contact of the hand at the lips is indicated by the asterisk.
therefore more suitable to use for Deaf persons outside the academic context. Figure 3.12 shows the DGS sign “question” (German: “Frage”) which is the noun form of the right sign in Figure 3.9 in SignWriting notation.

Note that all of the above annotation systems describe signs at a sublexical level and capture how a sign is performed, i.e. how it looks. On the other hand, glosses are base forms of a spoken language word, written in capital letters, which roughly correspond to the meaning of the sign. Since the same sign can have several meanings in different contexts, it can be transcribed differently depending on its context. In contrast to this, the term ID-gloss [Johnston 01] is used if one sign is always annotated with the same gloss, independent of its meaning in a specific context.

The corpora used in this thesis, the RWTH-PHOENIX-Weather 2014 Corpus and the Corpus-NGT, are annotated with ID-glosses. This means that if two signs look the same, they receive the same ID-gloss, although their meaning might differ depending on their context. Note however that the RWTH-PHOENIX-Weather 2014 Corpus was solely annotated for the use of statistical learning methods and not for linguistic purposes, so we refrained from creating a sign lexicon, which many linguists regard as a prerequisite for ID-glossing.

Sign variants are usually denoted by appending a common stem (or lemma) with a suffix, which simplifies the normalisation of variants for the use of machine translation.

Another advantage of using glosses instead of more fine-grained phonetic annotation systems is that the annotation work can be done faster and requires less training for the annotators, and thus larger corpora can be annotated with the same resources. Our group therefore chose to use gloss annotation when creating the RWTH-PHOENIX-Weather 2014 corpus.

However, gloss annotation also has disadvantages and limits. One difficulty arises with classifier predicates which are used to describe complex spatial phenomena such as size, shape, location, movement or other properties of an entity. The problem of annotating such classifier predicates using glosses is to map the complex three dimensional movements into a finite set of glosses.

Moreover, in sign languages, information is not only conveyed by the hands, but also by facial expression, eye gaze, torso or shoulder movements and by head movements such as head shakes. While some of this information can be marked by special tokens within the gloss system (for example by prefixing a gloss with “NEG-” to indicate negation expressed by a head shake), capturing the multimodal nature of sign languages in a one-dimensional stream of glosses will inevitably lead to some loss of information. For a critical discussion of gloss notation see [Pizzuto & Rossini 06]. Apart from machine translation, gloss notation also has disadvantages for sign language recognition compared to phonetic annotations, because without further information, a gloss cannot be broken down into smaller units similar to speech recognition where words are dissected into phonemes. See [Koller & Ney 13] for an approach to automatically annotate subunits based on our gloss corpus and a SignWriting dictionary.

In Section 5.2.3, we describe the annotation scheme used for the RWTH-PHOENIX-Weather 2014 corpus which is an enriched gloss notation including among others information about spatial locations, mouthings and the holding of a sign with one hand while signing another sign with the other hand.

3.3 Automatic Sign Language Recognition

In this section, we give a brief introduction to automatic sign language recognition. In this work, we do not work on the sign language recognition system itself but use the output of the recognition system as input and translate the gloss sequence of recognised signs into a text of a spoken language (see Figure 1.2 for the overall system architecture).

In isolated sign recognition, the goal is to recognise a single sign a signer has performed,
In the SignSpeak project, our group focused on continuous sign language recognition, where the goal is to recognize a sequence of signs continuously performed by a signer, recorded in a video.

An architecture overview of a continuous sign language recognition system can be seen in Figure 3.13. The architecture, proposed in [Dreuw & Rybach+ 07], is similar to the architecture of the statistical machine translation system in Figure 3.2, since both are based on Bayes decision rule.

The main difference between the systems lies in the different feature dimensionality and monotonicity constraints. The input features of the sign language recognition system $X_1^T$ are video frames over time, which leads to a large dimensionality, whereas the input to a machine translation system $f^T_1$ is a sequence of words in the source language, which results in a much smaller input dimensionality. The resulting features in sign language recognition are usually continuous (float values), while in sign language translation, discrete words are used. However, sign language recognition is monotonous, i.e. the glosses are recognized in the order in which they appear in the video sequence. On the other hand, in machine translation the order of the words in the source and target language may differ, especially in language pairs with different grammatical constructions. This leads to a larger complexity in machine translation, since the computer has to calculate the probability for different ways to reorder the source words. Consequently, in spite of a similar architecture, the implementation details between the two systems differ because of the different complexities.

The current recognition system models whole signs with a hidden Markov model in Bakis topology [Bakis 76], where each pair of consecutive states is modelled by a Gaussian mixture model with a globally pooled covariance matrix. The Bakis topology is depicted in Figure 3.14, it models the fact that a sign can be performed at different speeds. A rather slow performance is modelled by looping in the same state, a fast performance by skipping a state. The state transition probabilities also depicted in the figure thus model the probability of different signing speeds. Although our group did some research on modelling sign subunits [Koller & Ney+ 13], our main system models whole signs. The length of a sign and thus the number of HMM states are estimated from manual annotations of the time boundaries of individual signs.

An important aspect of the sign language recognition system is the feature analysis, i.e. the question what kind of features $x_1^T$ should be extracted from the raw video input $X_1^T$. 

![Diagram of continuous sign language recognition system](image-url)
As the research on sign language recognition in our group focused on hand-based features, the gloss annotation of the RWTH-PHOENIX-Weather corpus was also mainly focused on hand based signs, and often sign variants which differ only in non-manual aspects, such as mouthing or facial expression, were labelled identically. Thus, some information is lost in the recognition process. To mitigate this loss, we use a viseme recognition system based on an active appearance model as an additional input to our translation system (see Section 7.2). Active appearance models and viseme recognition are introduced in the next sections.

3.4 Active Appearance Models

To automatically detect facial expressions and mouthings, we apply the technique of active appearance models to track the contours of the face, eyes and mouth. The models were implemented by Thomas Hoyoux from the Intelligent and Interactive Systems, University of Innsbruck, our partner in the SignSpeak project (see Section 1.4). From the lower level features such as the position of the lip contours and eye corners, which are detected by the active appearance model, we derive higher level features such as horizontal and vertical mouth opening as well as eyebrow raise. These features will be used for two purposes: first to cluster signs with identical hand component based on their facial expression and mouthing, and second to detect the words pronounced by the signer and add this information to the translation process directly.

An active appearance model (AAM) is used to track key locations, also called landmarks, on the cheeks and chin outlines, the nose ridge and nose base, the eyelids and eye corners, the eyebrow outlines and the lip and mouth corners, as illustrated in Figure 3.15. We track these point features in the sign language videos in order to extract information about the mouth and eyebrows such as mouth openness and eyebrow raise. Table 3.2 lists the high-level features and the landmark points they are based on. Since the structure of the human face as described by a set of such point features exhibits a lot of variability due to changes in pose and expression, we base our tracking strategy on the deformable model registration method called active appearance models (AAMs).

Active appearance models, first proposed in [Edwards & Taylor+ 98] and further refined in [Matthews & Baker 04], belong to the family of deformable models for image interpretation. Such model-based methods recover the structure of an object displayed in an image by registering a deformable shape model of the object to the image data. Mathematically, the shape $s$ of an object
3 Preliminaries

Figure 3.15: Visualisation of facial annotations

Table 3.2: High-level facial features used for clustering and viseme recognition and the underlying lower-level point features (see Figure 3.15)

<table>
<thead>
<tr>
<th>Semantic description</th>
<th>Related points</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouth vertical openness</td>
<td>{18, 21, 24, 25, 26, 27}</td>
</tr>
<tr>
<td>mouth horizontal openness</td>
<td>{18, 21}</td>
</tr>
<tr>
<td>lower lip to chin distance</td>
<td>{26, 27, 32, 33}</td>
</tr>
<tr>
<td>upper lip to nose distance</td>
<td>{15, 16, 17, 18, 21, 24, 25}</td>
</tr>
<tr>
<td>left eyebrow state</td>
<td>{0, 1, 2, 6, 8}</td>
</tr>
<tr>
<td>right eyebrow state</td>
<td>{3, 4, 5, 10, 12}</td>
</tr>
<tr>
<td>gap between eyebrows</td>
<td>{2, 3}</td>
</tr>
</tbody>
</table>

is defined as the vector of coordinates of its \(v\) landmark points:

\[
s = (x_1, y_1, x_2, y_2, \ldots, x_v, y_v)^T
\]

, assuming that each landmark is a 2-dimensional point representing a salient point of the object such as the corner of an eye.

AAMs model shape deformation using a point density model (PDM), which is a parametric linear subspace model statistically trained with a principal component analysis (PCA) on a set of sample training shapes. The training examples have to be annotated by a human, resulting for example in the annotation in Figure 3.15 for the human face. In such a representation, any shape \(s\) of the deformable object can be expressed by the generative model as a base shape \(s_0\) plus a linear combination of \(n\) shape vectors \(s_i\):

\[
s = s_0 + \sum_{i=1}^{n} p_i s_i
\]

Adapting a PDM to a given image then is equivalent to determining suitable coefficient values \(p_i\) of the linear combination, i.e. the optimal PDM parameters. AAMs model the coupling between the PDM and the image data, predicting the PDM landmark locations on a given target image. The appearance model is itself a parametric linear subspace model, obtained by applying a PCA
to shape-normalised training images of the object of interest. This shape normalisation is done by warping every example image to a reference frame, which is usually performed by piecewise affine warping functions defined between each example shape and the base shape $s_0$ of the PDM. The generative appearance model is then used to express any object’s appearance $A(x)$ as a base appearance $A_0(x)$ combined with a linear combination of $m$ appearance images $A_i(x)$:

$$A(x) = A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) \quad \forall x \in R(s_0)$$

where $R(s_0)$ denotes the set of pixel locations within the region defined by the base shape $s_0$, i.e. the reference frame for the object’s appearance.

With the above two generative models and the application of the independent AAM definition presented in [Matthews & Baker 04], registration can be regarded as an image matching problem between the synthetic model image and the shape-normalised target image; the image fitting objective can be formulated as finding the parameters $p = (p_1, p_2, \ldots, p_n)^\top$ and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)^\top$ that minimise the following sum of squared differences:

$$\sum_{x \in R(s_0)} \left[ A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x) - I(W(x; p)) \right]^2$$

where $I$ is the target image and $W(x; p)$ is a (piecewise affine) warping function which projects a pixel location $x$ from the reference frame to the target image frame, depending on the PDM’s parameters $p$. The minimisation of this quantity is non-linear in the parameters $p$ and can be calculated iteratively by linear approximation, e.g. by applying the Gauss-Newton algorithm.

Several variants of this algorithm exist, usually differing in the way they model the linear approximation to derive the parameter update equation. In this work, we apply the efficient version of the simultaneous inverse-compositional AAM (SICAAM) presented in [Gross & Matthews+ 05]. This variant is especially robust to large variations in shape and appearance, which are typical when dealing with facial expressions in the context of sign language. Moreover, in order to cope with large off-plane head rotations, which happen when the signer looks to one side, which is common in sign language and can lead a 2D AAM to failure, we applied the refinement proposed in [Xiao & Baker+ 04]. In this work, a 3D PDM is estimated using a non-rigid structure-from-motion algorithm on the training shapes, and is then involved in the optimisation process which incorporates a regularisation term encouraging the 2D shape controlled by the 2D PDM to be a valid projection of the 3D PDM. Similar to the 2D PDM, the 3D PDM expresses any 3D shape $S$ as a 3D base shape $S_0$ plus a linear combination of $\bar{n}$ 3D shape vectors $S_i$:

$$S = S_0 + \sum_{i=1}^{\bar{n}} \bar{p}_i S_i$$

Notice that we also use the 3D PDM to calculate the high-level facial features described below.

The procedure for the production of the high-level facial features includes the following training phase:

1. Extrude the set of 2D training shape examples to 3D by means of the 3D PDM.
2. Remove global translations and rotations by aligning every extruded shape to the base shape $S_0$ of the 3D PDM.
3. Project the aligned extruded shapes to 2D and, for each shape, estimate local area-based measurements corresponding to the point features subsets given in Table 3.2.
4. For each point feature subset, store as the training output the minimum and maximum values of the corresponding local area-based measurements.

Extracting high-level facial features from the tracked lower-level point features is then done in the following way:

1. Extrude the registered shape and remove its global translation and rotation by means of the 3D PDM

2. Project the aligned extruded shape to 2D and, for each point features subset given in Table 3.2, estimate the corresponding local area-based measurement.

3. Normalise each local area-based measurement between 0.0 and 1.0 according to the minimum and maximum values obtained during training for the corresponding point features subset.

4. Each registered shape is then associated with a vector of $D$ (in our work $D = 7$) continuous values in the range $[0, 1]$, corresponding to our high-level facial features.

Figure 3.16 shows the application of an active appearance model to a video sequence of the RWTH-PHOENIX-Weather 2014 corpus. The model detects the position of the eyes, mouth and other landmarks on the face. The grid model in Figure 3.17b shows the recognised landmarks after a rotation into the frontal position. From these normalised points, horizontal and vertical mouth openness, the distances between chin and the lower lip as well as between the upper lip and nose are estimated based on the distance of corresponding points (for example, the distance between points 18 and 21 in Figure 3.15 is used to calculate the horizontal mouth openness). These measures are shown as a function over time in Figure 3.17c. The high-level features will be used in the clustering approach in Section 7.1 as well as for viseme recognition which is covered in the next section.

### 3.5 Viseme Recognition

In this section, we describe a viseme recognition system, also referred to as a “lip reading system”, which is used to recognise mouthings, i.e. words or syllables silently pronounced by the signer. The viseme recogniser itself was mainly developed by my colleague Oscar Koller (see [Koller & Ney + 15] for a recent development), but the alignment of glosses and translations to obtain possible mouthings from unlabelled data, the mapping of the pronunciation dictionary to the viseme level and the integration of the recognition results into the translation system are part of this work (see Section 7.2). A general survey of mouth modelling and mouthing recognition can be found in [Antonakos & Roussos + 15].

Since the RWTH-PHOENIX-Weather 2014 corpus used in this thesis (see Section 5.2) was mainly annotated for research on sign language recognition of hand-based features, mouthings have not been annotated for all signs in the corpus. To obtain training material for the viseme recogniser of the words the signer pronounces while signing, we align the glosses denoting the signs with their translation in the spoken language. We use the open-source toolkit GIZA++ to align each gloss to at most one word in the spoken language. However, not all signs are accompanied by mouthings. We therefore include a silence model representing no mouth movement and a garbage model for mouthing gestures which do not represent specific viseme sequences in the viseme recogniser. To train a viseme recogniser on the videos, we need a viseme transcription of the spoken words. We first use a pronunciation lexicon from the speech recognition system developed at our department trained on German to look up each German word which is aligned to a gloss.
3.5 Viseme Recognition

Figure 3.16: Active appearance models are used to track landmarks on the face (green grid lines) in a video sequence. The coloured bars indicate the degree of mouth openness (red; third bar), eye openness (blue; second, fourth bar) and eyebrow raise (yellow; first, fifth bar)

and to find its corresponding sequence of phonemes. The lexicon was generated using grapheme-to-phoneme conversion [Bisani & Ney 08]. As many phonemes cannot be visually distinguished (for example, the phonemes P and B differ only in the aspiration which is not visible) we further map the set of phonemes to a set of visemes, i.e. visually distinguishable phonemes. We follow the suggestion of [Weiss & Aschenbrenner 05] and map the set of phonemes to a set of 15 visemes. A list of the used visemes can be found in Table 3.3.

Statistics on the aligned gloss translation pairs allow us to exclude noisy alignments. Specifically, this is done by using an empirically set threshold of at least four occurrences per gloss translation pair and considering only translation alignments that represent at least 10% of all translations for a specific gloss. Gloss-translation alignments which do not meet these requirements are put into the garbage model. The above parameters have been obtained by optimisation on the development set.

We then train the state-of-the art speech recognition system RASR [Rybach & Gollan+ 09] using 15 viseme hidden Markov models (HMMs) and the garbage model, each containing three states with single Gaussian densities, a globally pooled covariance matrix and global time distortion penalties, i.e. Bakis HMM transition probabilities. Silence visemes are represented by an additional single state HMM. The models are fed with the seven high-level facial features. A lexicon defining possible pronunciation variants for each gloss is provided to the system. It is generated based on the statistics on the aligned gloss translation pairs. The system is initialised with a linear segmentation on the RWTH-PHOENIX-Weather 2014 data providing gloss time boundaries. The EM algorithm with Viterbi approximation iteratively accumulates the HMMs and uses them to re-estimate the state-frame alignment, while choosing the most likely pronunciation variants representing different sequences of visemes. This process can be considered as a form of weakly supervised clustering. After 10 iterations the algorithm converges to a stable optimum,
Figure 3.17: High-level feature extraction: features are extracted by tracking the face in the video (a), rotating the grid into frontal position (b), and measuring distances between the points (c).
### Table 3.3: Phoneme-viseme mapping. Phonemes in one row cannot be visually distinguished.

<table>
<thead>
<tr>
<th>Phoneme</th>
<th>Viseme</th>
<th>Example words</th>
<th>Phoneme transcriptions of examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>p, b</td>
<td>P</td>
<td>Pause, Bitte</td>
<td>paUz@, bit@</td>
</tr>
<tr>
<td>t, d, k, g</td>
<td>T</td>
<td>Tonne, Dach, König, Gier</td>
<td>tOn@, dax, k2niC, gi6</td>
</tr>
<tr>
<td>n, l</td>
<td>N</td>
<td>Nadel, Liebe</td>
<td>nad@l, lib@</td>
</tr>
<tr>
<td>m</td>
<td>M</td>
<td>Mutter</td>
<td>mUt6</td>
</tr>
<tr>
<td>f, v</td>
<td>F</td>
<td>Finder, Vase</td>
<td>fInd6, vaz@</td>
</tr>
<tr>
<td>s, z</td>
<td>S</td>
<td>Fass, Sein</td>
<td>fas, zaIn</td>
</tr>
<tr>
<td>S, Z, tS, dZ</td>
<td>Z</td>
<td>Schein, Garage, Tscheche</td>
<td>SaIn, gara:Z@, tSEC@</td>
</tr>
<tr>
<td>h, r, x, N</td>
<td>R</td>
<td>Hase, Reden, Dach, Wange</td>
<td>ha:z@, red@n, dax, vaN@</td>
</tr>
<tr>
<td>j, C</td>
<td>C</td>
<td>Junge, Wicht</td>
<td>jUN@, vICt</td>
</tr>
<tr>
<td>a:, a</td>
<td>A</td>
<td>Wagen, Watte</td>
<td>va:g@n, vat@</td>
</tr>
<tr>
<td>o:, O</td>
<td>O</td>
<td>Wolle, Wogen</td>
<td>vOl@, vO@n</td>
</tr>
<tr>
<td>u:, U</td>
<td>U</td>
<td>Buch, Runde</td>
<td>bu:x, rUnd@</td>
</tr>
<tr>
<td>@, 6</td>
<td>Q</td>
<td>Bitte, Weiher</td>
<td>bit@, va16</td>
</tr>
<tr>
<td>y6, Y, 2, 9</td>
<td>Y</td>
<td>Tür, Mütter, Goethe, Götter</td>
<td>ty6, mYt6, g2t©, g9t6</td>
</tr>
<tr>
<td>i6, I, e, E; E</td>
<td>E</td>
<td>Bier, Tisch, Weg, Räte, Menge</td>
<td>bi6.tIS, veg, Räte, Menge</td>
</tr>
</tbody>
</table>
Table 3.4: Viseme Error Rate (VER) and recall of viseme recogniser measured on 640 manual annotations.

<table>
<thead>
<tr>
<th></th>
<th>VER [%]</th>
<th>Recall [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>initial segmentation</td>
<td>40.5</td>
<td>82.5</td>
</tr>
<tr>
<td>10x EM-realignment</td>
<td>35.7</td>
<td>47.5</td>
</tr>
<tr>
<td>after RANSAC processing</td>
<td>32.2</td>
<td>45.5</td>
</tr>
</tbody>
</table>

yielding the hypothesised viseme sequences for each gloss. In order to remove outliers we chose the RANSAC algorithm [Fischler & Bolles 81] to further refine the state-frame alignment and hence the models. The RANSAC (random sample consensus) algorithm randomly chooses a subset of all samples (here: frames) and calculates the models based on this subset. If the resulting model leads to a high probability on most samples (the inliers) and to a low probability only on a few samples (the outliers), the model is used, else another random sample is drawn. The algorithm can robustly estimate model parameters and detect outliers if the relative number of outliers is not too high.

Table 3.4 shows the achieved performance of the viseme recogniser after training and optimisation. The Viseme Error Rate (VER) compares the hypothesised viseme sequence to 640 manually annotated mouthings. The recall is the fraction of samples which were used to train the models, the other samples were discarded as outliers.

Subsequently, the hypothesised viseme sequences are filtered by comparing them to the original GIZA++ alignment and estimating the relative error for a given gloss and viseme sequence. Viseme sequences that cause a high mismatch to the GIZA++ alignment are less likely to support the following translation step. We tested different error thresholds on the development set and obtained best results for a threshold of 30. Translation variants with a relative error above the threshold were removed, that is, no gloss variant was generated.

In Section 7.2, we combine the hand-based glosses and the output of the viseme recognition system in the translation system to deal with the multimodality of sign languages.

### 3.6 Summary

In this chapter, we reviewed the technology of statistical machine translation, which we want to apply to the translation of sign languages. We examined the particular features of sign languages, especially its multimodality, i.e. its ability to convey meaning simultaneously via several information channels (hand shape, movement, facial expressions, mouthing, eye gaze, etc.). To include mouthings as an additional feature into the decoding process, we use a viseme recognition system.

After a review of related work, in the following chapters we will enhance a statistical machine translation system to the task of translating from a sign language into a spoken language. We will examine a suitable annotation scheme for sign languages, improve the alignment procedures to work on sign language corpora which are rather small, and combine the output of a hand-based sign recognition system with a viseme recogniser to obtain better translation results.
In this chapter, we review related work in the field of sign language machine translation and sign language corpora. We also cover related work on machine translation using scarce data resources as well as work dealing with the multimodal aspects of sign languages. We discuss the state of the art in the field and highlight where it is lacking.

4.1 European Research Projects

While this work was written as part of the European FP-7 project SignSpeak (see Section 1.4), several other European research projects also worked on computer-based sign language processing, each differing in focus and approach to sign languages.

In the ViSiCAST project (2000–2002, see [Elliott & Glauert 00]) and the subsequent eSIGN project (2002–2004, see [Elliott & Glauert+ 08]), the aim was to develop a spoken language to sign language translation system translating from a text in the spoken language (English) to an avatar animation in the sign languages (British Sign Language, Dutch Sign Language and German Sign Language). An interlingua-like representation called “discourse representation structures” was used to translate into the different sign languages. The purpose of the system is to convert the text of dynamic websites into a video of an avatar signing the content of the website. Since the websites are constantly changing, it is not feasible to produce videos with human interpreters translating the text into a sign language.

In the DICTA-SIGN project (2009–2011, [Efthimiou & Fontinea+ 10]), the aim was to make online communications more accessible to Deaf sign language users by the use of Web 2.0 technologies. Webcam-based sign recognition and avatar animations were developed in the project. One application of the developed technology is the anonymisation of signing by replacing the human signer with an animated avatar.

In contrast to the other projects, the SignSpeak project realised the full translation pipeline from a sign language video to a spoken language text.

4.2 Sign Language Corpora

To this day, sign languages are under-resourced when it comes to corpora suitable for statistical methods such as statistical sign language recognition or machine translation. One of the main reasons for the lack of sufficiently large corpora is the fact that sign languages do not have an official written form (see Section 3.2.4), and each corpus annotation project has to decide on or develop an annotation scheme. These annotation schemata are often explicitly designed for the research
questions of the project in which they were annotated. For example, if a research programme examines the role of eye gaze in signing, the gaze direction is added to the annotation, while it is usually not annotated. As a consequence, corpora from different research projects usually differ in the applied annotation scheme and are difficult to compare or combine. Other reasons for smaller sizes of sign language corpora are the fact that the scientific community working on sign languages is much smaller than the community working on spoken languages, and the fact that the annotation time for a video-based sign language corpus is considerably longer than for a spoken language corpus, mostly due to the complex annotation of multimodal information.

In this section, we review existing sign language corpora with respect to their utility for statistical sign language translation. We thus restrict ourselves to corpora which contain translations into a spoken language. For a more general discussion of sign language corpora which can be used for sign language recognition, see [Forster & Schmidt+12]. An updated list of sign language corpora used for linguistic research was created by the University of Hamburg, Germany.\footnote{www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/sl-corpora.html}

In recent years, projects for collecting national sign language corpora were founded in many countries.

For American Sign Language, [Athitsos & Neidle+08] created a sign language video collection called the ASL lexicon video data set. The videos contain sequences of signs, but the main aim of the dataset is to create a video lexicon to look up signs and their meaning using vision-based recognition technologies. A later development is the American Sign Language Linguistic Research Project at Boston University,\footnote{http://www.bu.edu/asllrp} which collected and annotated a large corpus of 1 888 sign language utterances [Neidle & Vogler 12]. Besides ID-glosses for each sign, parts of the corpus are annotated with additional information such as head movements, topic/focus of the sentence or role shift. While translations seem to exist, they are not available via the data access interface at the moment.

The British Sign Language Corpus Project [Schembri & Fenlon+13] consists of an annotation of ID-glosses and a free translation into English. Studies on lexical frequency and phonological variations led to an annotation of individual signs of the video for parts of the corpus. Corpus statistics are only given for the annotation of individual signs for a lexical variation study (7 332 signs) and a phonological study (6 330 signs), but not for the annotations of gloss sentences and their free translations.

For Australian Sign Language, [Johnston 08] presents the Auslan corpus which has no specific domain but consists of interviews, conversations and narratives. The roughly 150 hours of video footage have been annotated with ID-glosses, and great care was taken to obtain consistent lexical annotations. Additionally, a free translation into English is provided.

[Bungeroth & Stein+08] present the ATIS Sign Language Corpus, a corpus in the domain of air travel information. The corpus consists of 595 sentences in Irish Sign Language, German Sign Language and South African Sign Language as well as German and English. While interesting because of its multiple languages, the corpus size is rather small for data-driven and statistical approaches.

For Spanish Sign Language, [López & San-Segundo+10] present a corpus in the domain of identity document and driving licence renewal. The corpus consists of 4 080 sentences, annotated by ID-glosses and a Spanish translation. Their work on this corpus is presented in the next section.

As the above description of related sign language corpora shows, only a few corpora were created for the application of statistical machine learning methods, but rather for linguistic or sociological studies. Apart from the ATIS corpus and the corpus on identity document renewal, the corpora do not have a concise domain and are too small for statistical methods.
Our work focuses on two corpora, the RWTH-PHOENIX-Weather Corpus and the Corpus-NGT. The RWTH-PHOENIX-Weather 2014 Corpus is one of the largest sign language corpora available and has a concise domain of weather forecasting, making statistical methods applicable. For a detailed description of the corpus see Section 5.2. While we selected a subset of the Corpus-NGT which is suitable for machine translation, the broader domain of Deaf related discussions and the smaller size led to less than satisfactory results. We therefore focused on the RWTH-PHOENIX-Weather 2014 Corpus and only present some interesting findings with respect to the Corpus-NGT on the use of independent annotation of the left and right hand and of head shake annotation in Section 7.3.1.

4.3 Sign Language Machine Translation

In this section, we review sign language machine translation systems. Note that systems translating from a sign language into a spoken language are structurally very different from systems translating from a spoken language into a sign language. While the former require a sign language recognition system to recognise signs in a video, the latter need an animation system to display an animated virtual signer called an “avatar”, which performs the signs. The different system architectures are depicted in Figure 4.1, an example of an avatar is given in Figure 4.2. While our work focuses on the translation from a sign language into a spoken language, in Section 7.1 we detect sign variants with differing facial expressions which could be used to improve the avatar animation. We do however not develop or apply an avatar system.

[Veale & Conway + 98] present an early rule-based system called Zardoz which translates English text into ASL, ISL (Irish Sign Language) and JSL (Japanese Sign Language). The system implements a blackboard control structure and transforms the English text into an interlingua (see Section 1.2), from which the signs in a specific sign language are generated. No experimental results are reported.
Matt Huenerfauth [Huenerfauth 04a, Huenerfauth 06] developed a multipath approach to machine translation from English to ASL to cover different aspects of sign languages such as classifier predicates and the arrangement of spatial verbs as well as the token space. The translation system architecture consists of different modules, and for each sentence the system decides whether to produce a classifier predicate or use a transfer approach to translate the other sentences. The classifier predicate generation module uses an interlingua approach where the objects described in the source sentence are visualised in the three dimensional signing space.

For the restricted domain of weather reports, which we also address with the RWTH-PHENIX-Weather Corpus in Section 5.2, [Grieve-Smith 99] creates an interlingua system translating forecasts from English to ASL. The system consists of a lexical analyser, a parser, a transfer module and a generation module. The written weather forecasts taken from the Albuquerque National Weather Service use a highly formulaic register with a simple syntactic structure which can be fully parsed, and the standardised information such as sky, precipitation, wind and heat is stored in an interlingua, from which the sign language utterances are generated. While this method seems feasible for the register of the above corpus, it is not applicable to corpora with more complex sentence structures even in the restricted domain of weather forecasts, e.g. when the meteorologist speaks freely and uses a colloquial register, let alone in corpora with a broader domain.

Only a few research groups apply data-driven methods to the task of sign language machine translation. One of the reasons is the lack of sufficiently large corpora. The National Centre for Language Technology at Dublin City University developed MATREX, a data-driven machine translation system which combines example-based and statistical machine translation approaches [Stroppa & Way 06]. In [Morrissey 08], [Morrissey 11] and [Morrissey & Way 13], the MATREX system is applied to the task of sign language machine translation, both on the multilingual ATIS corpus (see Section 4.2) and a medical dialogue data corpus in Irish Sign Language and English. Translation results are evaluated both by automatic measures and a human evaluation judging the avatar output of the system.

Closely related to our work is the work by Rubén San-Segundo and his group on Spanish Sign Language (Lengua de signos o señas española, LSE). In [López & San-Segundo 10], they use the Moses phrase-based decoder to translate from LSE to Spanish and examine different alignment merge heuristics (see [Stein & Schmidt 12, Chapter 5.1] for our work on this topic). In the next section, we also present their work on the data scarcity issues in the sign language domain.

Summarising, most research groups apply linguistically motivated rule-based methods to the task of sign language machine translation. The groups which do apply data-driven methods usually work with glosses, which do not take into account the multimodal nature of sign language. Moreover, current data-driven research is usually hampered by small corpus sizes. In the following section, we review related work on how to use the scarce training resources available efficiently.
4.4 Scarce Language Resources

The statistical methods described in Chapter 3 and applied in this thesis depend on a sufficient amount of training data for a reliable estimation of probabilities. This holds especially true for the automatic alignment which is based on the IBM-models and obtained with the GIZA++ toolkit (see Section 3.1.2). When applying GIZA++ to a sign language corpus, which usually consists only of a few hundred or thousand utterances, the alignment quality is rather poor.

In the following, we review works that aim at handling data scarcity problems which are either due to scarce training resources or to morphologically rich languages which have many surface forms, i.e. many inflected forms of a common base form, resulting in poor probability estimates for each form.

[Nießen & Ney 04] construct a hierarchical lexicon model for the case of German to English translation. They use a morphological analyser on each word to obtain the lemma form and information about its inflection. The difference in morphology between German and English is automatically learnt using a maximum entropy framework.

For the task of German to German Sign Language, [Stein & Bungeroth+ 06] use a morphological analyser and apply three heuristics: gender information contained in German suffixes are removed, noun compounds are split, and function words which are not used in DGS, e.g. articles, are removed.

When translating from German into another language, splitting noun compounds can lead to improvements, as the above publication shows. If however one translates into German, this method cannot be applied, because the resulting text would contain split noun compounds, which is not consistent with German orthography. However, one can use compound splitting only during alignment training to improve the alignment between the individual nouns and their translation in the other language and restore the original orthography afterwards. [Popović & Stein+ 06] implemented this technique for the task of English to German translation. In [Stein & Schmidt+ 12, Chapter 6.3], we applied the technique to the translation of DGS to German. For more information and examples, see Section 5.4. In Section 6.1 we further refine this method by resolving ambiguous noun compound splits.

For Spanish Sign Language, the group of Rubén San-Segundo worked on the data scarcity issues of sign language corpora. In [López-Ludeña & San-Segundo+ 11], they use a rule-based reduction of morphological complexity by mapping inflected Spanish words to a manually created base-form referred to as “tags”. In [López-Ludeña & San-Segundo+ 12], they automated the process of categorization. In [López-Ludeña & San-Segundo+ 13], they handle the data scarcity problem by introducing morphological variants for some names, adjectives and verbs and common expressions. In comparison, in Sections 6.1 and 6.2 we automatically derive mappings between inflected words and their corresponding glosses by using a morphosyntactic analysis and a synonym database to improve the word alignment process.

[Schwenk 08] uses semi-supervised training for French to English machine translation by translating large amounts of monolingual data using a translation system trained on a smaller amount of bilingual data. We apply and extend this method on a text collection of German weather forecasts in Section 6.3.

In summary, many methods exist to apply a morpho-syntactic analysis to improve the translation of language pairs with scarce resources. Few approaches have been taken to apply these methods to the task of sign language machine translation. We will use morpho-syntactic methods, a synonym database and semi-supervised training to improve the translation quality of a sign language machine translation system in Chapter 6.
4.5 Multimodality

In this section, we review works dealing with the multimodality of sign languages. We focus on manual features as the main modality and on facial expressions and mouthings as an important additional modality. The importance of facial features and mouthings for detecting sign variants in German Sign Language is shown in [Von Agris & Knorr + 08].

[Massó & Badia 10] try to integrate mouthings and mouth gestures in a Catalan to Catalan Sign Language system by using the Moses factored models. However, their small corpus of only 153 training sentence leads to issues in parameter optimisation.

Apart from the above work on machine translation, most work focuses on integrating several modalities in the recognition process. [Forster & Oberdörfer + 13] compare different approaches to combine several modalities in recognition, either on the feature level, the hidden Markov state level, or on the gloss level. [Ma & Gao + 00] similarly combine a hand gesture recogniser and a lip motion recogniser using a parallel hidden Markov model architecture. [Yang & Lee 11] integrate manual and non-manual features using conditional random fields and support vector machines.

The main difference between the above approaches and this work is not only the handling of multimodality in recognition versus in translation. While the above approaches need a ground truth distinguishing gloss variants based on all modalities, our work can start from a hand-based gloss annotation and automatically infers gloss variants based on their translation in the spoken language. Thus, our approach can extend an existing hand-based system to new modalities without any additional annotation effort.

[Aran & Burger + 09] uses a belief-based sequential fusion approach to merge the manual information and head motion in order to distinguish sign variants on a set of 29 signs. If a sign cannot be detected with a high confidence based on the manual features alone, a second stage classification is applied based on the non-manual information. Similarly, we use a clustering technique to distinguish between sign variants in Section 7.1.

[Huenerfauth 04b] designed a detailed representation of ASL which consists of multiple channels, is hierarchically structured and coordinated over time. Since he focuses on translation into ASL, the representation is mainly used for avatar animation.

This work focuses on the exploitation of the multimodality of sign languages when translating from the sign language into a spoken language. For the RWTH-PHOENIX-Weather, we basically use a tuple-based annotation scheme to integrate multiple modalities, which can be linearised, see Section 5.2.3. Moreover, we combine different modalities in decoding by combining manual glosses with mouthing information in Section 7.2. Lastly, in Section 7.3, we also apply our translation system to the Corpus-NGT which is annotated using multiple channels, featuring a channel for each hand and for head shakes. We thus present different approaches to handle multimodality in sign language machine translation.
Sign languages have no official writing system. For the linguistic research of sign languages, different annotation systems have been developed (see Section 3.2.4), and often the annotation differs from corpus to corpus.

In this chapter, we want to analyse the impact of the annotation scheme on the translation quality. To this aim, we examine the RWTH-PHOENIX-Weather Corpus, which was annotated at our group as part of the SignSpeak project. We particularly discuss how the multimodal nature of sign language is captured in the annotation scheme by enriching the gloss annotation with additional information such as mouthings, negations, repetitions or the location of an action in the signing space. We found that especially mouthings and locations in the signing space are important aspects of the annotation necessary for a good translation quality. To our knowledge, this is the first work comparing the impact of different annotation schemes on machine translation.

The chapter is structured as follows: First, we discuss the annotation software ELAN, which was used to annotate the RWTH-PHOENIX-Weather Corpus, in Section 5.1. In Section 5.2, we introduce the RWTH-PHOENIX-Weather Corpus itself and describe the annotation scheme in detail. To obtain a bilingual corpus suitable for machine translation, in this work we developed an efficient way to annotate the spoken language side of the corpus by employing an automatic speech recognition system in Section 5.3. In Section 5.4, we present the machine translation setup used in our experiments. We then show experimental results on the RWTH-PHOENIX-Weather Corpus in Section 5.5. We examine the effects of the sign language recognition system and the sign language translation system on the overall performance, both with automatic metrics and with a human evaluation by sign language experts. Moreover, we evaluate the impact of the annotation scheme on the translation quality by performing contrastive experiments on different schema variants.

5.1 Annotation Tools

The annotation of video-based sign language corpora is a very laborious and time-consuming task. An annotation tool must support the playback and framewise navigation through a video to annotate time boundaries of individual signs and utterances. In the sign language research community, the following two tools are well-established: ELAN [Wittenburg & Brugman + 06] and ILEX [Hanke 02]. Both tools have roughly the same functionality when it comes to annotating a corpus on the gloss level. Since our research partner in the SignSpeak project from the Radboud University Nijmegen uses ELAN to annotate their Dutch sign language corpus (Corpus-NGT, see Section 7.3.1), we also chose to use the tool for the annotation of the RWTH-PHOENIX-Weather
Corpus. Moreover, ELAN features multiple annotation tiers, so the different aspects expressed
simultaneously while signing can be annotated on several tiers. Screenshots of the ELAN toolkit
with examples from the RWTH-PHOENIX-Weather and the Corpus-NGT which are examined
in this thesis can be seen in Figure 5.1. The figure also shows the multi-tier annotation used in
the two corpora. In the RWTH-PHOENIX-Weather Corpus, we used one tier to mark whole sign
language utterances, one tier for individual gloss time boundaries, and one tier for the German
translation. In the Corpus-NGT, the group of Onno Crasborn used two sets of tiers, each set for
one of the two signers who engage in a dialogue. For each signer, left and right hand are annotated
individually, head movements such as head shakes or nods are annotated on an additional tier,
and one tier contains the translation into Dutch.

5.2 The RWTH-PHOENIX-Weather Corpus

The RWTH-PHOENIX-Weather Corpus is a video-based large vocabulary corpus of German Sign
Language. Following the philosophy of using real-world data instead of laboratory data produced
for academic research only, the corpus consists of weather forecasts recorded from German public
TV. The TV station “Phoenix”\(^1\) regularly broadcasts the major public news programmes with an
additional interpretation into German Sign Language, using an overlay window which shows the
interpreter. The interpreters, who are usually bilinguals raised in families with Deaf relatives, do
not receive a transcript of the news in advance, but have to interpret the news live while listening
to the announcer speaking. This form of live interpretation is a very difficult task, as the news
programme is very compact, and information is conveyed very fast. E.g. the weather forecasts of
the next three days are presented in a limited time span of only 1.5 minutes.

The sign language interpretation was manually annotated by our group using an enriched gloss
system. Further, the text spoken by the announcer was semi-automatically transcribed using a
state-of-the-art automatic speech recognition system (see Section 5.3). The RWTH-PHOENIX-
Weather 2014 Corpus is one of the largest annotated sign language corpora available today, and
although it is small when compared to spoken language corpora, the choice of the compact domain
of weather forecasting makes it suitable for statistical methods.

5.2.1 Motivation

Many sign language corpora are recorded and annotated by the linguistic community to study
specific linguistic characteristics of sign languages. They are usually unsuitable to train statistical
models, because the selection or construction of the sentences contained in the corpus lead to
relative frequencies of the signs which are different from regular sign language usage. For example,
a corpus which examines spatial verbs would consist of a set of sentences each containing such a
verb. Thus, the relative frequencies of these verbs would all be equal and much higher than their
usage in regular signing. Consequently, to obtain a corpus suitable to train a statistical model,
one has to use data consisting of signs and sentences which occur with a natural frequency.

The German public broadcast channel “Phoenix” features live interpretation of the main
evening broadcast news into German Sign Language. The RWTH-PHOENIX-Weather Corpus
consists of whole show segments and thus preserves the natural usage of the language in the
domain. The original plan of our group was to use the whole news programmes for the corpus.
However, as news programmes cover a wide variety of topics which change on a daily basis, we did
not expect to be able to annotate enough data with the given budget and time constraints and
consequently restricted ourselves to the weather forecasts which are featured at the end of each
\(^1\)www.phoenix.de
5.2 The RWTH-PHOENIX-Weather Corpus

(a) The RWTH-PHOENIX-Weather Corpus in ELAN. The annotation is done on three tiers:
1. gloss utterance (“default”),
2. glosses with individual time boundaries (“wort”),
3. German translation sentences (“translation”)

(b) The Corpus-NGT in ELAN. Four tiers are annotated for each of the two signers:
1. left hand glosses with individual time boundaries (“GlosL”),
2. right hand glosses with individual time boundaries (“GlosR”),
3. Dutch translation sentence (“SignSpeakZin”),
4. head shakes with time boundaries (“head movement”)

Figure 5.1: Corpus annotation in ELAN of the RWTH-PHOENIX-Weather and the Corpus-NGT
show. The domain of weather forecasting is rather limited in its vocabulary but still rich in sign language phenomena such as classifiers and use of the signing space. For an example screenshot of the videos used in the RWTH-PHOENIX-Weather Corpus see Figure 5.2.

Admittedly, the language modelled by a system which is trained on the RWTH-PHOENIX-Weather Corpus is restricted to the domain of weather forecasting and thus only covers a small fraction of the whole language. Moreover, as [Filhol & Tannier 14] correctly point out, the use of interpreter’s videos has two disadvantages. First, the live interpretation process can result in a bias towards the word order and peculiarities of the spoken source language. Second, due to the real time constraints, the interpreter can make more translation errors than signers in a setup in which several takes can be recorded until a good quality is obtained. We mitigate the effect of interpretation errors by providing a second German reference translation which is closer to the signed utterances (see Section 5.2.4). Despite its downsides, the corpus also has several strong points. The domain of weather forecasting is particularly rich in spatial usages, as weather phenomena are depicted on an imaginary map in front of the signer. Facial expressions are used to indicate the degree of the weather phenomena, and mouthings express geographical names such as cities, rivers or areas. The corpus stresses the importance of non-manual features, and we think that it is useful to develop and analyse statistical recognition and translation systems and study the effect of the annotation scheme on the translation quality. After a remark on the different versions of the corpus, we discuss the annotation scheme used for the corpus.

5.2.2 Corpus Versions

The recording and annotation of the RWTH-PHOENIX-Weather Corpus was a continuous effort of our group. Consequently, different versions were created in the course of time. Based on earlier works in our group of [Bungeroth & Stein+ 06] and [Stein & Forster+ 10], we published two versions of the corpus, and we refer to these versions by the year of their publications, namely RWTH-PHOENIX-Weather 2012, published in [Forster & Schmidt+ 12] and RWTH-PHOENIX-Weather 2014, published in [Forster & Schmidt+ 14]. The two versions mainly differ in size, as more shows were recorded, annotated and added to the corpus over time. To keep the results in this work consistent, we present most of our experiments on the latest version, the RWTH-PHOENIX-Weather 2014 Corpus. The only exception is Section 5.5.1 where we report on a human evaluation done at the end of the SignSpeak project. Since it takes considerable effort
5.2 The RWTH-PHOENIX-Weather Corpus

Table 5.1: Annotation Scheme of the RWTH-PHOENIX-Weather 2014 Corpus (glosses translated into English)

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Example</th>
<th>number of running glosses</th>
</tr>
</thead>
<tbody>
<tr>
<td>gloss in capital letters</td>
<td>AS-ALWAYS</td>
<td>87941</td>
</tr>
<tr>
<td>finger spelling split by +</td>
<td>A+G+N+E+S</td>
<td>40</td>
</tr>
<tr>
<td>compound glosses split by +</td>
<td>V+LAND</td>
<td>1954</td>
</tr>
<tr>
<td>numbers in written form</td>
<td>SEVEN instead of 7</td>
<td>3248</td>
</tr>
<tr>
<td>pointing gestures</td>
<td>IX</td>
<td>2091</td>
</tr>
<tr>
<td>extended repetitions</td>
<td>SUN++</td>
<td>876</td>
</tr>
<tr>
<td>classifier signs</td>
<td>cl-COME</td>
<td>953</td>
</tr>
<tr>
<td>left hand only signs</td>
<td>lh-SUN</td>
<td>571</td>
</tr>
<tr>
<td>signs negated by head shake</td>
<td>neg-WIND</td>
<td>109</td>
</tr>
<tr>
<td>signs negated by the alpha rule</td>
<td>negalp-MUST</td>
<td>46</td>
</tr>
<tr>
<td>localisation</td>
<td>loc-RAIN</td>
<td>2842</td>
</tr>
<tr>
<td>additional mouthing</td>
<td>GLOSS-(mb:hill)</td>
<td>4471</td>
</tr>
<tr>
<td>additional facial expression</td>
<td>GLOSS-(mk:strong)</td>
<td>1333</td>
</tr>
<tr>
<td>additional localisation</td>
<td>GLOSS-(loc:alps)</td>
<td>6703</td>
</tr>
<tr>
<td>additional object of sign</td>
<td>GLOSS-(obj:cloud)</td>
<td>383</td>
</tr>
</tbody>
</table>

and the necessity of bilingual experts to conduct a human evaluation, we refrained from repeating the evaluation and instead present the original results performed on the earlier version RWTH-PHOENIX-Weather 2012. For general statements which are true for all versions, we use the general term “RWTH-PHOENIX-Weather Corpus” without a version number.

5.2.3 Annotation Scheme

The RWTH-PHOENIX-Weather Corpus is annotated with ID-glosses. In contrast to the Corpus-NGT, the annotation scheme does not contain individual annotation tiers for each hand but assumes that the sign is performed with the dominant (which in the case of the corpus is the right hand since all interpreters are right-handed) hand of the signer or with both hands simultaneously. Additional information about mouthings, the object of a sign, or the retention of a sign with the left hand while continuing to sign with the right hand are annotated by enriching these glosses.

The annotation scheme is shown in Table 5.1. The table also shows the number of occurrences in the RWTH-PHOENIX-Weather 2014 Corpus for each kind of annotation. Note that the different kinds of additional information schemata (lowest group of rows in the table) are not mutually exclusive but can be combined. For example, in one instance the interpreter signed the sign for “Schleswig-Holstein”, a federal state of Germany, which was annotated as SCH+H+STEIN- (mbschleswig-holstein), which means that she used finger spelling SCH and H and the sign for stone (“Stein”), combined with the mouthing of the word “Schleswig Holstein” (“mb”, is short for “Mundbild”, German for “mouthing”). The annotation can be considered a sparse tuple annotation, as the additional information (mouthing, facial expression, localisation and object) can
5 Annotation Schemata for Machine Translation

Table 5.2: Example for the application of preprocessing rules
(Upper group of rows: original German glosses, lower group of rows: English translation for reference)

<table>
<thead>
<tr>
<th>glosses</th>
<th>example sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>loc-REGION-(loc:westraum) SECHZEHN BIS ZWEIZWANZIG GRAD</td>
</tr>
<tr>
<td>preprocessed German</td>
<td>REGION WESTRAUM $number {16} BIS $number {22} GRAD</td>
</tr>
<tr>
<td>raw</td>
<td>loc-REGION-(loc:westregion) SIXTEEN TO TWENTYTWO DEGREE</td>
</tr>
<tr>
<td>preprocessed English</td>
<td>REGION WESTREGION $number {16} TO $number {22} DEGREE</td>
</tr>
</tbody>
</table>

be present or absent for a sign. E.g. the above sign could be stored as the tuple

\[
\begin{bmatrix}
SCH \\
H \\
STEIN \\
mb : schleswig – holstein \\
mk : - \\
loc : - \\
obj : -
\end{bmatrix}
\]

in which the gloss and a mouthing annotation is stored, but facial expression, localisation and the object of the sign are empty.

The raw annotation contains a lot of variants for the same sign because of different values for the additional information, leading to a rather low type/token ratio of 12.4 and a singleton rate of 66%. Thus, a suitable preprocessing scheme has to be applied to reduce the number of variants.

As a baseline setup, we use the following preprocessing rules:

* for additional information, create new glosses
* convert written numbers into digits
* use placeholders for numbers, ordinal numbers, weekdays and months

In Table 5.2, we explain the applied preprocessing rules using an example sentence. We will examine the effect of different preprocessing schemata on the translation quality in Section 5.5.3.

In many statistical machine translation systems, special placeholders are used for numbers written in digits, because they are usually left untranslated, and in the standard procedure without placeholders, phrases would be extracted for the specific number only, which does not generalise well. The open-source machine translation system Moses [Koehn & Hoang 07] uses the category tag @num@ to replace numbers. The open-source translation system Jane developed at our group applied in this work uses categories preceded by a dollar sign, e.g. $number for numbers. In this work, we extended the category system to include numbers (e.g. “1500”) , ordinal numbers (e.g. “3.”), weekdays (e.g. “Montag”, German for “Monday”), months (e.g. “Januar”, German for “January”). These categories occur quite often in the domain of weather forecasting, and the resulting phrases containing category symbols generalise from specific numbers, weekdays, etc.

5.2.4 Corpus Setup

One of the aims of the “SignSpeak” project (see Section 1.4) was to develop a vision-based technology for translating continuous sign language into a spoken language text. The RWTH-PHOENIX-Weather corpus was therefore developed for the creation and evaluation of both automatic sign
language recognition and translation systems. We defined two test sets, one containing only utterances of the most frequent interpreter (single signer), the other containing utterances by several signers (multi signer). Note that the identity of the signer has hardly any effect on the translation system, but it has a strong effect on the error rates in sign language recognition, which highly depends on the appearance, the signing style and regional signing variants of the interpreters. In this work, we use the single signer setup only in the pipeline experiments of combining the recognition and translation system (Section 5.5.1), because recognition of the multi signer setup was too challenging at the time of our experiments. In the other sections we use the multi signer setup, because it contains a larger test set.

The corpus statistics of the preprocessed RWTH-PHOENIX-Weather 2014 Corpus for the single signer and multi signer setup can be found in Table 5.3. The corpus consists of a total of 8834 sentences. A few explanations on the terms used in the table: *singleton* are words which have been seen only once in training. *Out-of-vocabulary words* (OOVs) are words which occur in the development or test set but not in the training set.

The test set of the bilingual corpus was defined in [Forster & Schmidt + 14] and the selection of the test segments was mainly done with sign language recognition in mind. The development set was randomly selected from the training corpus by splitting off one sixth of the training sentences.

<table>
<thead>
<tr>
<th>Table 5.3: Statistics of the RWTH-PHOENIX-Weather 2014 Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glosses</strong></td>
</tr>
<tr>
<td>Training:</td>
</tr>
<tr>
<td>Sentences</td>
</tr>
<tr>
<td>Running Words</td>
</tr>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>Singletons</td>
</tr>
<tr>
<td>Development:</td>
</tr>
<tr>
<td>Sentences</td>
</tr>
<tr>
<td>Running Words</td>
</tr>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>OOVs (running words)</td>
</tr>
<tr>
<td>Test (multi signer):</td>
</tr>
<tr>
<td>Sentences</td>
</tr>
<tr>
<td>Running Words</td>
</tr>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>OOVs (running words)</td>
</tr>
<tr>
<td>Test (single signer):</td>
</tr>
<tr>
<td>Sentences</td>
</tr>
<tr>
<td>Running Words</td>
</tr>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>OOVs (running words)</td>
</tr>
</tbody>
</table>

A sentence in one language can have multiple translations in other languages. Since in machine translation automatic scores are calculated by comparing the system output to a human reference translation, the allowable variability in translations is ideally captured by providing more than one translation. Moreover, sometimes the sign language interpreters did not convey all information spoken by the announcer, mostly for time constraints, and thus there partly is an information mismatch in the corpus. We therefore created an additional reference for the RWTH-PHOENIX-Weather 2014 corpus by translating the glosses back into German. Consequently, the additional reference is closer to the information contained in the glosses. For an example of a mismatch
5 Annotation Schemata for Machine Translation

<table>
<thead>
<tr>
<th>German</th>
<th>Morgen reichen die Temperaturen ...bis 17 Grad im Breisgau.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glosses</td>
<td>MORGEN ... MAXIMAL 17 GRAD region suedwest raum</td>
</tr>
<tr>
<td>2. Ref.</td>
<td>Morgen reichen die Temperaturen ...bis 17 Grad im Süden.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English</th>
<th>Tomorrow, the temperatures reach ...17 degrees in the Breisgau.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glosses</td>
<td>TOMORROW MAXIMUM 17 DEGREE region south-west region</td>
</tr>
<tr>
<td>2. Ref.</td>
<td>Tomorrow, the temperatures reach ...17 degrees in the south west.</td>
</tr>
</tbody>
</table>

(a) An example of a mismatch between the spoken language and the sign language in the RWTH-PHOENIX-Weather Corpus (sentence shortened). While the announcer uses “Breisgau”, an area in the south-west of Germany, the interpreter simply signs “south-west region”. The second reference follows the glosses more closely. (English translation added below the bar for reference.)

(b) The sign for southwest region. Note that the sign has to be seen from the perspective of the signer. Consequently, to indicate “west”, she moves her flat hand to the bottom left.

Figure 5.3: An example of a mismatch between the spoken announcement and the sign language interpretation in the RWTH-PHOENIX-Weather Corpus is shown in (a). The sign for “southwest region” performed by the interpreter instead of the region name “Breisgau” can be seen in (b).

between the glosses and the spoken text and the creation of an additional reference, see Figure 5.3.

5.3 Efficient Annotation of Spoken Language Side

The sign language interpretation of the RWTH-PHOENIX-Weather corpus was annotated at our group by a Deaf and a hard-of-hearing expert, both of them have German Sign Language as their mother tongue. To create a bilingual corpus, we also needed to annotate the text spoken by the announcer and align the spoken language side and the sign language side on the sentence level.

For the RWTH-PHOENIX-Weather 2014, the annotation was done as part of this work. Speech recognition transcripts were kindly provided by my colleagues Jens Forster and Zoltan Tüske, who applied our open-source automatic speech recognition system RASR [Rybach & Gollan+ 09] which was trained on German broadcast data from the French research programme QUAERO [Sundermeyer & Nußbaum-Thom+ 11]. To create complete transcripts, the speech recognition output has to be further refined: punctuation has to be added, all recognition errors have to be corrected, and as the speech recognition system was a lower case system, the words have to be put into the correct case. The overall processing pipeline can be seen in Figure 5.4.
After obtaining the correct transcript by manually correcting the output of the automatic system pipeline, we can calculate the errors made by each of the automatic methods, see Table 5.4.

The error rate of the speech recognition system was roughly 4.5%, which means that instead of having to type the whole text, one has to correct roughly every 22nd word of the automatic transcript. Moreover, since many recognition errors consisted of a wrong word ending or a wrong separation of words, these mistakes could be manually fixed quickly, often without having to pause the playback of the audio.

For punctuation prediction, we follow the approach of [Peitz & Freitag 11] and use a mono- tonic phrase-based system to insert the punctuation. The system is trained on pseudo-bilingual data by using a text corpus from the domain of weather forecasting, the DWD dataset (see Section 6.3). On the target side, the text including punctuation is used, but on the source side, the punctuation is removed. Instead of using GIZA++, the word alignment can be generated directly when removing the punctuation from the source side, because the words on source and target side are identical apart from the punctuations. The error rate of automatic punctuation prediction was rather high. However, we found that often the system predicted the existence of a full stop or a comma correctly, but confused the two.

For automatic case conversion, the frequent case approach was used, which means that for each word, the case which occurs in a training corpus (again the DWD dataset) most often is used. Moreover, the first word of each sentence is capitalised. The frequent case approach leads to only a few casing errors on the corpus (2.3%), because words with ambiguous casing like nominalised verbs (which are capitalised in German) are rather uncommon in the domain.

By using the automatic methods to create a transcript which contains punctuations as well as upper and lower case spelling, the amount of manual work could be dramatically reduced in comparison to the manual typing of the whole text or the manual addition of all punctuations and correct case spellings. Even though the total error rate of 12.3% which includes punctuation and casing errors seems to be rather high, the wrong punctuation and casing can be quickly corrected
Table 5.4: Errors of the automatic processing pipeline

<table>
<thead>
<tr>
<th>System</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatic Speech Recognition</td>
<td>4.5%</td>
</tr>
<tr>
<td>Automatic Punctuation Prediction</td>
<td>5.5%</td>
</tr>
<tr>
<td>Automatic Case Converter</td>
<td>2.3%</td>
</tr>
<tr>
<td>Total Word Error Rate</td>
<td>12.3%</td>
</tr>
</tbody>
</table>

on the fly while listening to the audio file in real-time.

5.4 Machine Translation System Setup

In this section, we describe the statistical machine translation system which was used in our experiments in the following chapters. To train a statistical machine translation system, many decisions have to be taken, e.g. the choice of an alignment heuristic, of the approach taken (phrase-based, hierarchical, etc.), of a suitable development set size, of an optimisation criterion, etc. In [Stein & Schmidt+ 12], we optimised our statistical translation system for the case of sign language machine translation and arrived at the following system setup.

We use the open-source statistical machine translation system Jane [Vilar & Stein+ 12] to train three systems: a phrase-based system, a hierarchical phrase-based system, and a hierarchical system with shallow-1 grammar (see Section 3.1 for a description of the different approaches).

The word alignment of the corpus has been trained using GIZA++. We extract alignments in both directions and use the final-and-grow-mono heuristic to merge both alignment directions. The final-and-grow-mono heuristic, introduced in [Och & Ney 03], starts from the intersection of the two alignments. It adds alignment points from the union alignment when both the source and the target word are unaligned (“final and”). Then it “grows” the obtained alignment iteratively by adding alignment points which are part of the union alignment if they are horizontally or vertically adjacent to points already in the alignment. Alignment blocks are avoided by allowing only vertical or horizontal extensions at a time.

From the word aligned corpus, phrases with a maximum length of 6 words on the source side and 12 words on the target side are extracted. For language modelling, a 4-gram language model with Kneser-Ney smoothing and interpolation was used.

Category symbols for numbers, ordinal numbers, weekdays and months were generated using simple regular expression rules. In the phrase extraction phase, these symbols are extracted instead of concrete numbers, weekdays, etc. This leads to generalised phrases. Phrases with inconsistent categories, e.g. a different number of categories on the source and target side, which can result from a mismatch of information in the two languages, are not extracted. When a phrase contains more than one category, the GIZA++ word alignment is used to know which category symbols in the source and target phrase match. When translating the development or test set, the concrete value of a category symbol is passed through the decoder, and in the end, the category symbol is again replaced by its concrete value.

We use the morphological analyser “Morphisto” [Zielinski & Simon 08] to split German noun compounds, align the glosses to the split corpus using GIZA++ and then merge the compounds again to obtain an improved alignment between the glosses and the German text. When splitting the noun compounds, the splitting points are stored, and in a post-processing step they are merged again. Consequently, the alignment points have to be merged as well. See Figure 5.5 for...
Morphisto does not only split noun compounds but also performs a morphological analysis to each word and returns its base form, i.e. dictionary form, as well as grammatical information such as gender, number and case for nouns or person and tense for verbs. We use the base form returned by Morphisto to further improve the alignment between glosses and words. Since we know that glosses are usually also base forms, we increase the probability of aligning glosses and words with identical base forms. This is done by adding parallel sentences consisting of only one gloss and the identical spoken language word to the bilingual corpus. We further give these gloss-word pairs a higher sentence weight in the GIZA++ training and roughly optimise the weight on the development corpus by varying the parameter. Best results were obtained for a weight of 10.

5.5 Experimental Results

In the following sections, we present experimental results on the RWTH-PHOENIX-Weather corpus. Since it was the goal of the SignSpeak project to create a full sign language video to spoken language text translation system, we examined the pipeline of a sign language recognition system and a sign language translation system in Section 5.5.1. Since the recognition error rate on the corpus is still above 30%, we conduct the other experiments on the gloss transcriptions.

In Section 5.5.2, we compare three statistical machine translation approaches, the phrase-based, hierarchical phrase-based approaches and a hierarchical approach with shallow-1 grammar and see that on the corpus at hand, the shallow-1 grammar leads to best results.

In Section 5.5.3, we examine the importance of the additional information provided in the gloss annotation and find that especially mouthing and locations in signing space are important aspects for the corpus at hand.

5.5.1 Effect of Recognition and Translation Errors

The overall goal of a sign language-to-spoken language translation system is to translate a video of a person signing into the text of a spoken language with the same meaning. This process is usually implemented as a two-step process [Bauer & Nießen+ 99]. First, an automatic sign recognition system (see Section 3.3) is used to extract the signs in some form. These signs, which in this system are represented in the form of a gloss sequence, are then translated into a spoken language utterance using a machine translation system.

In the SignSpeak project, our group realised the full pipeline of sign language recognition and translation by combining our sign language recognition system (which is based on our open-source speech recognition system RASR [Rybach & Gollan+ 09] and which was mainly developed by Philippe Dreuw, Jens Forster, Yannik Gweth and Oscar Koller, see Section 3.3) and our sign language translation system (which is based on the open-source translation system Jane [Vilar & Stein+ 12] and which was mainly developed by Daniel Stein and the author of this thesis).

Here we present the results obtained towards the end of the SignSpeak project on the RWTH-PHOENIX-Weather 2012 Corpus. Unfortunately, we could not repeat the human evaluation on the RWTH-PHOENIX-Weather 2014 Corpus because of the time and effort of the sign language experts whom we could not come by outside of the project context.

Since the overall performance of the sign language recognition system was not satisfactory in 2012 when the SignSpeak project ended, having a recognition error rate of 54.4%, we also performed experiments simulating a better recognition performance by artificially decreasing the language model perplexity in the recognition system. Thus, several hypotheses with different recognition error rates were produced. For the experiment, we used the best translation system developed in the course of the project. The effect of the recognition error rate on the translation
5 Annotation Schemata for Machine Translation

(a) Baseline alignment.

(b) Split alignment.

(c) Crunched alignment.
5.5 Experimental Results

Figure 5.5: Example of the alignment crunching technique, taken from the RWTH-PHOENIX-Weather 2014 corpus, on the sentence “Tagsüber bleibt die Wolkendecke in Süddeutschland geschlossen.” (Engl.: “During the course of the day, the cloud cover in southern Germany remains dense.”). The word “Wolkendecke” (“cloud cover”) is a singleton, but “Wolke” (“cloud”) is of course well-known and “Decke” is known from “Schneedecke” (“snow cover”). Thus, in (a) the alignment has errors, but for compound split German in (b) the quality is much better. After crunching the alignment in (c), the alignment structure matches the original German sentence.

Table 5.5: The effect of the sign language recognition quality on the machine translation results, measured on the RWTH-PHOENIX-Weather 2012.

<table>
<thead>
<tr>
<th>Input</th>
<th>PPL</th>
<th>Recognition WER [%]</th>
<th>Translation BLEU [%]</th>
<th>Translation TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript</td>
<td></td>
<td>0</td>
<td>39.2</td>
<td>53.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>34.6</td>
<td>57.0</td>
</tr>
<tr>
<td>Language models</td>
<td>3.1</td>
<td>15.4</td>
<td>23.1</td>
<td>67.7</td>
</tr>
<tr>
<td>with reduced</td>
<td>6.4</td>
<td>30.8</td>
<td>19.8</td>
<td>71.0</td>
</tr>
<tr>
<td>Output perplexity</td>
<td>15.0</td>
<td>48.7</td>
<td>16.6</td>
<td>75.8</td>
</tr>
<tr>
<td>Regular language model</td>
<td>34.9</td>
<td>54.4</td>
<td>15.0</td>
<td>75.1</td>
</tr>
</tbody>
</table>

The effect of the sign language recognition quality on the machine translation results is shown in Table 5.5 and Figure 5.6. The graph shows that with an increasing recognition error rate, the translation quality drops dramatically. Note that the current recognition system does not recognise the additional information but only the raw glosses, which already leads to a loss in information (first group of rows in the table, dashed lines in the graph). We later introduce a viseme recognition system to recognise mouthings and include this information in the translation process in Section 7.2.

In addition to the automatic evaluation, we also performed a human evaluation in cooperation with the “Deaf and Sign Language Research Team” (DESIRE) at RWTH Aachen University. A total of 9 evaluators (6 Deaf and 3 hearing) performed the evaluation. Since the human evaluation of sign language videos is rather time consuming, we restricted the study to 18 video segments.

For the evaluation, each evaluator was sitting in front of a PC. First, the video of the signed utterance was presented to the evaluator. After the first play of the video, the recognised text was shown as well. The evaluators could replay the video as often as they liked. Then, they could evaluate the video.
The evaluation was conducted according to two criteria as is common in machine translation research (see [Koehn 10, Chapter 8] for details).

- The adequacy measures whether the translated text conveys the meaning of the signed video.
- The fluency measures whether the translations are grammatically correct and fluent.

For each evaluation, the evaluators applied a score from 1 ("poor") to 5 ("excellent"). We selected some videos which were translated by running the pipeline of recognition and translation ("recognition"), some videos in which the human annotated glosses were translated ("transcription"), and some videos in which the original text of the announcer was shown to the evaluator ("reference"). The adequacy and fluency scores can be found in Figure 5.7. The comparison of the three groups indicates two points. First, when comparing the adequacy scores, the difference between the recognition setup and the transcription setup is larger than the difference between the transcription setup and the reference score. This indicates that the errors introduced by the sign language recognition system affect the preservation of meaning more strongly than the errors introduced by the translation system. Second, the fluency scores of the recognition and the transcription setup are very similar. This indicates that the translation system produces output with an average fluency irrespective of the quality of the input. The human evaluators especially gave lower fluency scores when the translation contained grammatical mistakes. To summarise, the errors of the recognition system severely affect the adequacy of the output, whereas the errors introduced by the translation system mainly affect the fluency and grammatical correctness of the output.

To examine whether the automatic measures used in machine translation correlate to human evaluations, we compare the TER scores of the individual videos to the average adequacy and fluency scores applied by the human evaluators. We refrain from calculating the correlations using BLEU scores, as they are defined on the corpus, not the sentence level, though sentence level variants exist. The results can be seen in Figure 5.8. An ideal Pearson correlation, which
5.5 Experimental Results

Figure 5.7: Human Evaluation of the translation results of the RWTH-PHOENIX-Weather 2012 Corpus: adequacy and fluency scores are indicated by the bars, the vertical lines indicate the variance between the evaluators.

Figure 5.8: Correlation between the automatic metric TER and human evaluation scores measures linear correlation, is -1.0, as lower TER scores are better, but the optimum human evaluation score was 5 in the range of 1-5. The correlation between TER and accuracy is -0.62, which indicates only a rough correlation. The correlation between TER and fluency is -0.80, which indicates a stronger correlation.

5.5.2 Baseline Results RWTH-PHOENIX-Weather 2014

With the system setup and the different approaches described in Section 5.4, we obtain the following results on the RWTH-PHOENIX-Weather 2014 corpus, see Table 5.6. We can see that the use of categories, of Morphisto to split noun compounds and of the optimisation method similar to n-fold cross-validation improve the translation quality on all three approaches, the phrase-based, the hierarchical phrase-based, and the hierarchical phrase-based system with shallow-1 grammar. Among the three, the shallow-1 grammar method turned out to give best performance of 29.8% BLEU and 62.9% TER on the RWTH-PHOENIX-Weather 2014 Corpus. In the following experiments, we will therefore use this setup.
Table 5.6: Translation results on the RWTH-PHOENIX-Weather 2014 multisigner test set using different approaches and methods.

<table>
<thead>
<tr>
<th></th>
<th>Phrase-Based</th>
<th>Hierarchical</th>
<th>Shallow-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>TER [%]</td>
<td>BLEU [%]</td>
</tr>
<tr>
<td>Baseline</td>
<td>27.0</td>
<td>65.6</td>
<td>24.4</td>
</tr>
<tr>
<td>+ Categories</td>
<td>27.6</td>
<td>65.1</td>
<td>27.8</td>
</tr>
<tr>
<td>+ Morphiosto</td>
<td>28.8</td>
<td>63.7</td>
<td>28.3</td>
</tr>
<tr>
<td>+ n-fold cross-validation</td>
<td>29.3</td>
<td>63.5</td>
<td>28.7</td>
</tr>
</tbody>
</table>

Table 5.7: Translation quality of RWTH-PHOENIX-Weather 2014 depending on the use of additional information

<table>
<thead>
<tr>
<th></th>
<th>Dev BLEU [%]</th>
<th>TER [%]</th>
<th>Test BLEU [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw glosses</td>
<td>17.1</td>
<td>76.4</td>
<td>22.6</td>
<td>73.7</td>
</tr>
<tr>
<td>no location &amp; mouthing</td>
<td>19.8</td>
<td>74.3</td>
<td>23.1</td>
<td>72.8</td>
</tr>
<tr>
<td>no mouthing</td>
<td>20.7</td>
<td>74.6</td>
<td>26.2</td>
<td>68.0</td>
</tr>
<tr>
<td>no location</td>
<td>20.8</td>
<td>74.1</td>
<td>26.7</td>
<td>66.8</td>
</tr>
<tr>
<td>all additional information</td>
<td>21.0</td>
<td>73.9</td>
<td><strong>29.8</strong></td>
<td><strong>62.9</strong></td>
</tr>
</tbody>
</table>

5.5.3 Effect of Annotation on Translation

In this section, we want to examine the effect of additional context information which go beyond the manual features usually expressed in the gloss notation. In the RWTH-PHOENIX-Weather corpus, this information is added to the glosses using brackets, e.g. the notation L-(mb:lausitz,lh:region) indicates that the signer shows an L-shaped hand and pronounces the word “Lausitz” (a region in Germany), while his left hand indicates the region in the signing space. First, we examined the importance of the most common information contexts “loc”, indicating what location in the signing space an action refers to, and “mb” (mouthing, German: “Mundbild”) indicating what word the signer silently pronounced with his lips while performing the sign. In the current system, we use all of the information by creating additional glosses. Thus, the above gloss would be preprocessed as “L LAUSITZ REGION”. Table 5.7 shows the effect on the translation quality if we use only the raw glosses without any additional information, if we remove only location, mouthing, both, or if we use all additional information. The result shows that especially the mouthing information and the location information are crucial, and that removing this information greatly degrades the system performance. Other forms of additional information such as facial expressions, the holding of a sign with the left hand while signing a new sign with the right hand and the object of a sign were only of minor significance with regards to the translation quality.

We also experimented with creating special glosses by marking the glosses with the context they come from, e.g. processing the above example as “L mb-lausitz lh-region”. However, all such trials led to worse results than the simple baseline system to add the information to the gloss string as ”L LAUSITZ REGION".
5.6 Summary and Contributions

While we see the potential in the future to automatically generate some of this additional information in the recognition process, e.g. the recognition of mouthing, other information such as the location highly depends on knowledge-based models such as the knowledge about European geography or remembering the content of previous sentences. This information, while easily producible by the human annotator, is more difficult to be generated by a computer. Future annotation schemes should therefore focus on those aspects of the glosses which can be produced in the recognition phase. For German Sign Language, special focus should be put on mouthing, which turned out to convey a lot of important information.

5.6 Summary and Contributions

In this chapter, we presented the RWTH-PHOENIX-Weather 2014 Corpus, a video-based large vocabulary corpus of German Sign Language with translations in spoken German. The glosses of the corpus were annotated by Uwe Zelle and Kathi Kullrich-Zelle under the guidance of Daniel Stein and Jens Forster. The author efficiently annotated the spoken language side by using speech recognition system output and a punctuation and casing recovery and manually correcting the errors. The corpus was published at the LREC Conference for further research in the scientific community [Forster & Schmidt+ 14].

For the SignSpeak system pipeline of sign language recognition and translation, we examined how the errors introduced by each component affect the overall system performance and found that errors introduced by the recognition system distort the meaning more strongly than errors of the translation system, but that errors in the translation system mainly affect the fluency, i.e. the grammaticality of the output. The sign language recognition results were produced by Jens Forster, the sign language translation results were produced by the author. Moreover, the author designed and guided the human evaluation, which was performed by several sign language interpreters and deaf experts in collaboration with the “Deaf and Sign Language Research Team” (DESIRE) at RWTH Aachen University. The website with which we performed the evaluation and which played the videos and asked for the evaluation scores was implemented by Thomas Hoyoux from the University of Innsbruck, one of our partners in the SignSpeak Project.

Next, we compared different translation approaches and found that a shallow-1 grammar system including categories, noun compound splitting for a better word alignment and an optimisation using n-fold cross validation leads to the best translation quality on the RWTH-PHOENIX-Weather 2014 Corpus. The experiments on the current corpus were performed solely by the author.

With this system, the author examined the importance of the additional information for sign language machine translation and found that the most important additional information is mouthing, followed by locations in the signing space. This led us to the idea to include the mouthing information into our translation system by using active appearance models and a viseme recognition system in Chapter 7.
Statistical methods need sufficient amounts of training data to model a probability distribution of the language at hand. For spoken languages, bilingual corpora used for the training of machine translation systems are in the range of a few million sentences. For example, the Europarl corpus [Koehn 05] German-English Release v7 consists of 1.9 million sentences. What’s more, monolingual corpora for language model training have sizes of up to a few billion words. For example, the English Gigaword corpus [Graff & Cieri 03] consists of 1.8 billion running words.

Since sign languages lack a common writing system, annotated sign language corpora are mostly created in the academic field of linguistics and are currently much smaller (in the range of a few thousand sentences, see Section 4.2). This lack of sufficient training data often leads to a poor automatic alignment between the annotated signs and their corresponding words and phrases in the spoken language. This issue is aggravated by the fact that sign languages, which are expressed by moving the hands in front of the body, have a morphology that is totally different from spoken languages, which are produced by sequences of sounds and which for example use word inflections by changing the word ending.

We improve the automatic alignment with two methods: first, we apply a morphosyntactic analysis on the spoken language words to bridge the differences in morphology to find corresponding signs and phrases (Section 6.1). Moreover, we use a synonym database to find correspondences where gloss and the spoken word are synonymous (Section 6.2). As bilingual sign language corpora are rather small, we try to improve the translation quality by using a monolingual corpus of in-domain data, i.e. weather reports, first by training a larger language model, but also by translating the monolingual text corpus into glosses and using the resulting bilingual corpus for training (Section 6.3).

### 6.1 Improving Word Alignments Using Morphosyntactic Analysis

As described in Section 5.4, we use the morphological analyser “Morphisto” to split noun compounds on the spoken German side. As noun compounds are usually signed by performing the signs of the individual nouns in sequence, the gloss annotation consists of two individual glosses. E.g., the German noun compound “Wolkenlücke” (cloud gap) is translated into “WOLKE LOCH” (CLOUD HOLE). After splitting noun compounds, the individual words can be aligned with a higher confidence by the system.

Morphisto is not a probabilistic system. This means that if a word has more than one analysis, the analyser returns a list of all analyses in an arbitrary order and does not estimate which analysis is most likely. Multiple analyses can occur for two reasons. First, a compound
can often be split at different points, leading to different constituent words. Second, as words sometimes have the same surface form in different grammatical cases, a given surface form can have multiple analyses. Many of these analyses are grammatically correct, but do not make sense. For example, Morphisto decomposes the word “Grenzwetterlage” (English: “marginal weather conditions”, meaning “weather conditions at the border of two air masses”) into six different splits, see Table 6.1. These six split surface forms each correspond to multiple grammatical cases, resulting in a total of 24 analyses, which can be seen in Table 6.2.

In the original approach in [Stein & Schmidt+ 12, Chapter 5.2], we used a heuristic which selects an analysis with the minimum number of split words. In case of multiple analyses, the heuristic simply selects the first one (we refer to this method as the “first minimal” heuristic in the result tables.) In the above example, the heuristic selects the nonsensical “grenzen wettern Lage” (“to border/ to rant / situation”, underlined in Table 6.2). To improve this analysis, we train a probability model for the parts of speech (POS) sequences and choose the analysis with the minimal number of words and the most probable POS sequence. In the example, the correct split “Grenze Wetter Lage” (“border weather situation”) is derived (marked bold in the table).

Note that not all words have ambiguous analyses. Table 6.3 shows the statistics on the number of analyses of the RWTH-PHOENIX-Weather 2014 Corpus vocabulary. While only 12.4% of the words have only one analysis, most words (77.1%) have several analyses, but only one split base form corresponding to several grammatical cases.

Since we do not have data of noun compounds annotated with correct POS tags, we use the Morphisto analyses of words having a single analysis or a unique surface form as training data, ruling out the cases with a possibly incorrect surface form. Since the words are usually split into one to three constituents, we do not decompose the probability of the POS sequence further but model the probability of the whole POS sequence. The model is trained according to the maximum likelihood criterion, which leads to relative frequencies of the POS sequences seen in the training samples.
6.1 Improving Word Alignments Using Morphosyntactic Analysis

Table 6.2: An example of a morphosyntactic analysis of the word “Grenzwetterlage” (“border weather situation”). The original heuristic chooses the first analysis with the shortest length (underlined, “bordering ranting situation”). The approach proposed here chooses the analysis with the most probable parts-of-speech sequence (bold, correct split). The word “Wetter” can also mean “a betting person”. The analysis represents this as “wetten er” (V NN SUFF), a verb with a noun suffix.

<table>
<thead>
<tr>
<th>split base form</th>
<th>parts of speech, case, number</th>
<th>#words</th>
</tr>
</thead>
<tbody>
<tr>
<td>grenzen wetten er Lage</td>
<td>V V NN SUFF +NN Fem Gen Sg</td>
<td>4</td>
</tr>
<tr>
<td>grenzen wetten er Lage</td>
<td>V V NN SUFF +NN Fem Nom Sg</td>
<td>4</td>
</tr>
<tr>
<td>grenzen wetten er Lage</td>
<td>V V NN SUFF +NN Fem Dat Sg</td>
<td>4</td>
</tr>
<tr>
<td>grenzen wetten er Lage</td>
<td>V V NN SUFF +NN Fem Akk Sg</td>
<td>4</td>
</tr>
<tr>
<td>grenzen wetten Lage</td>
<td>V V +NN Fem Nom Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen wetten Lage</td>
<td>V V +NN Fem Gen Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen wetten Lage</td>
<td>V V +NN Fem Dat Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen wetten Lage</td>
<td>V V +NN Fem Akk Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze wettern Lage</td>
<td>NN V +NN Fem Gen Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze wettern Lage</td>
<td>NN V +NN Fem Nom Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze wettern Lage</td>
<td>NN V +NN Fem Dat Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze wettern Lage</td>
<td>NN V +NN Fem Akk Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen Wetter Lage</td>
<td>V NN +NN Fem Gen Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen Wetter Lage</td>
<td>V NN +NN Fem Nom Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen Wetter Lage</td>
<td>V NN +NN Fem Dat Sg</td>
<td>3</td>
</tr>
<tr>
<td>grenzen Wetter Lage</td>
<td>V NN +NN Fem Akk Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze wetten er Lage</td>
<td>NN V NN SUFF +NN Fem Gen Sg</td>
<td>4</td>
</tr>
<tr>
<td>Grenze wetten er Lage</td>
<td>NN V NN SUFF +NN Fem Nom Sg</td>
<td>4</td>
</tr>
<tr>
<td>Grenze wetten er Lage</td>
<td>NN V NN SUFF +NN Fem Dat Sg</td>
<td>4</td>
</tr>
<tr>
<td>Grenze wetten er Lage</td>
<td>NN V NN SUFF +NN Fem Akk Sg</td>
<td>4</td>
</tr>
<tr>
<td><strong>Grenze Wetter Lage</strong></td>
<td><strong>NN NN +NN Fem Gen Sg</strong></td>
<td><strong>3</strong></td>
</tr>
<tr>
<td>Grenze Wetter Lage</td>
<td>NN NN +NN Fem Nom Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze Wetter Lage</td>
<td>NN NN +NN Fem Dat Sg</td>
<td>3</td>
</tr>
<tr>
<td>Grenze Wetter Lage</td>
<td>NN NN +NN Fem Akk Sg</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 6.3: Ambiguity of the Morphisto analysis of the RWTH-PHOENIX-Weather 2014 Corpus vocabulary consisting of 2,860 words

<table>
<thead>
<tr>
<th># analyses</th>
<th># surface forms</th>
<th>counts</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>355</td>
<td>12.4%</td>
</tr>
<tr>
<td>&gt;1</td>
<td>1</td>
<td>2,205</td>
<td>77.1%</td>
</tr>
<tr>
<td>&gt;1</td>
<td>&gt;1</td>
<td>300</td>
<td>10.5%</td>
</tr>
<tr>
<td>total</td>
<td>2,860</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>
Table 6.4: Experimental results of the improved Morphisto analysis

<table>
<thead>
<tr>
<th></th>
<th>Development BLEU [%]</th>
<th>Development TER [%]</th>
<th>Test BLEU [%]</th>
<th>Test TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>no Morphisto</td>
<td>20.3</td>
<td>74.2</td>
<td>27.1</td>
<td>66.0</td>
</tr>
<tr>
<td>first minimal</td>
<td>21.0</td>
<td>73.9</td>
<td>29.8</td>
<td>62.9</td>
</tr>
<tr>
<td>maximum likely POS</td>
<td>21.9</td>
<td>70.6</td>
<td>31.0</td>
<td>61.3</td>
</tr>
</tbody>
</table>

Table 6.5: Example sentence from the RWTH-PHOENIX-Weather 2014 Corpus. The words “Westen” (west) and “Wetter” (weather) are correctly analysed by the maximum likely POS sequence approach, but incorrectly by the first minimal heuristic (as “wesen”, to rot, and “wettern”, to scold).

<table>
<thead>
<tr>
<th>Processing</th>
<th>Example sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td>im Westen und Nordwesten beruhigt sich später das Wetter.</td>
</tr>
<tr>
<td>first minimal</td>
<td>im wesen und Nord Westen beruhigen er späť das wettern.</td>
</tr>
<tr>
<td>maximum likely POS</td>
<td>im Westen und Nord Westen beruhigen er späť das Wetter.</td>
</tr>
<tr>
<td>English</td>
<td>In the west and north-west the weather later calms down.</td>
</tr>
</tbody>
</table>

With the analyses obtained by this approach, we calculate the word alignment using GIZA++ and retrain the machine translation system. In the result table, we refer to this method as the “maximum likely POS” heuristic. The translation results can be seen in Table 6.4. There is a consistent improvement on both the development and the test set. Note that the first line of the table “no Morphisto” differs from the result in Table 5.6 as it uses no Morphisto parsing, but n-fold cross validation. An example of the improved analysis can be seen in Table 6.5. We can see that the lack of sufficient training data leads to suboptimal alignments, and that this lack can at least in part be remedied with a morphological analysis.

6.2 Using Synonym Data to Improve Alignments

In the last section, we improved the alignment of glosses and words which only differ in inflection or compounding. There are however many cases in the corpus in which glosses and words
are not identical, but synonymous. For these cases, a morphological analysis does not lead to improvements. In this section, we want to improve the alignment of such pairs by using a synonym database. Some examples of the case that a gloss and its corresponding spoken word are synonymous can be found in Table 6.7. One reason for the large number of such pairs is the fact that at least for the case of German Sign Language and spoken German, the spoken language uses more lexical variation, whereas in German Sign Language one lexical item can be modified e.g. by changing the accompanying facial expression, the speed or mode of signing, etc. We can also see this difference in lexical variation by comparing the gloss vocabulary and the spoken language vocabulary: in the RWTH-PHOENIX-Weather 2014 Corpus training data, the vocabulary of the glosses is 1 371, the spoken language vocabulary is 2 860.

We use a German synonym database which is based on the Leipzig Corpora Collection ("Projekt Deutscher Wortschatz")\(^1\) which features semantic relations for 130 000 words [Quasthoff & Goldhahn\(^+\) 13]. The database was accessed using the freely available python API libleipzig-python.\(^2\)

We integrated the synonym information by adding sentence pairs containing glosses and their corresponding synonymous spoken words to the corpus on which we train the word alignment using GIZA++. The data is only used in the alignment phase but not for phrase extraction. After retraining the translation system with the extracted alignment, we obtain the following translation results: see Table 6.6. The results show substantial improvements both in BLEU and TER, indicating the validity of the method. An example of the improvements can be found in Figure 6.1. In the example, the gloss-word pairs STELLENWEISE, mitunter ("sporadically", "occasionally") and VERSCHWINDEN, lichten ("to vanish", "to thin out") are synonymous and were aligned by the procedure including the synonyms.

### 6.3 Additional Monolingual Data and Semi-Supervised Training

As annotated sign language corpora are rather small, we want to improve our translation system by employing monolingual data, which is easier to obtain. Matching to the domain of the RWTH-PHOENIX-Weather Corpus, we collected written German texts in the domain of weather forecasting by subscribing to a newsletter of the German Meteorological Service ("Deutscher Wetterdienst", or DWD for short). The corpus statistics of the DWD Corpus can be found in Table 6.8.

In a first step, we added the DWD data to the translation system by simply training an additional language model on the data and training the system using both the original and the new language model with one scaling factor for each model.

In a second step, we generated additional bilingual training data by training an inverse translation system, i.e. a system translating from Spoken German into glosses, on the

\(^1\)http://wortschatz.uni-leipzig.de
\(^2\)https://github.com/lehmannro/libleipzig-python
Table 6.7: Left: examples of glosses and their corresponding words in the German language text. The gloss-word pairs are synonyms.
Right: English translation of left table for reference.

<table>
<thead>
<tr>
<th>Gloss</th>
<th>German</th>
<th>Gloss</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACHTUNG</td>
<td>Vorsicht</td>
<td>ATTENTION</td>
<td>caution</td>
</tr>
<tr>
<td>BAUER</td>
<td>Landwirt</td>
<td>FARMER</td>
<td>agriculturalist</td>
</tr>
<tr>
<td>BESONDERS</td>
<td>vorwiegend</td>
<td>ESPECIALLY</td>
<td>predominantly</td>
</tr>
<tr>
<td>BRAUCHEN</td>
<td>benötigen</td>
<td>NEED</td>
<td>require</td>
</tr>
<tr>
<td>DESWEGEN</td>
<td>daher</td>
<td>THEREFORE</td>
<td>hence</td>
</tr>
<tr>
<td>LOCH</td>
<td>Lücke</td>
<td>HOLE</td>
<td>gap</td>
</tr>
<tr>
<td>NEBEL</td>
<td>Dunst</td>
<td>MIST</td>
<td>haze</td>
</tr>
<tr>
<td>TROPFEN</td>
<td>nieseln</td>
<td>TRICKLE</td>
<td>drizzle</td>
</tr>
<tr>
<td>SCHRANK</td>
<td>Truhe</td>
<td>CUPBOARD</td>
<td>chest</td>
</tr>
<tr>
<td>VORSTELLEN</td>
<td>präsentieren</td>
<td>PRESENT</td>
<td>present</td>
</tr>
<tr>
<td>KLAR</td>
<td>ungetrübt</td>
<td>CLEAR</td>
<td>unclouded</td>
</tr>
<tr>
<td>PASSEN</td>
<td>entsprechen</td>
<td>MATCH</td>
<td>correspond</td>
</tr>
<tr>
<td>SCHAFFEN</td>
<td>erreichen</td>
<td>ACCOMPLISH</td>
<td>achieve</td>
</tr>
<tr>
<td>SEHEN</td>
<td>schauen</td>
<td>SEE</td>
<td>look</td>
</tr>
<tr>
<td>VERSCHWINDEN</td>
<td>lichten</td>
<td>VANISH</td>
<td>clear up</td>
</tr>
</tbody>
</table>

Table 6.8: Corpus statistics for German Meteorological Service (“Deutscher Wetterdienst”) newsletter

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Newsletters</td>
<td>1 202</td>
</tr>
<tr>
<td>Sentences</td>
<td>42 190</td>
</tr>
<tr>
<td>Running Words</td>
<td>668 540</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>3 206</td>
</tr>
</tbody>
</table>
6.3 Additional Monolingual Data and Semi-Supervised Training

(a) Baseline alignment generated with GIZA++. The alignment quality is quite poor.

(b) Alignment generated with GIZA++ using a Morphisto-split target side.

Figure 6.1: Example of an improved alignment by using a synonym database. The left sentence has quite a poor alignment quality. On the right, the alignments of STELLENWEISE to mitunter and of VERSCHWINDEN to lichten are recognised, because the words are found in the synonym data base.
Figure 6.2: Graphical representation of semi-supervised training. The monolingual text data is translated using an inverse machine translation system (dashed arrow). The monolingual data and its translation are then used as additional training data for a new machine translation system.

The RWTH-PHOENIX-Weather 2014 corpus, translating the monolingual data into glosses and using the data obtained thus as additional bilingual data. A visual illustration of the semi-supervised training method is given in Figure 6.2. The idea of this method goes back to [Schwenk 08].

However, when simply translating the monolingual data without any additional preparation, the system performance deteriorated, and in a setup where we trained one scaling factor for the original phrases and one for the additional phrases produced by the semi-supervised method, the new phrases got a weight of zero during parameter optimisation. The main reason was that the monolingual data contained a lot of unknown words which could not be translated by the inverse translation system. Although the texts of the DWD corpus come from the same domain as the RWTH-PHOENIX-Weather corpus, the style in which the texts are written and the words used are slightly different. When using the corpus in our experiments, we therefore faced the problem of many out-of-vocabulary words, that is, words occurring in the DWD texts but not occurring in the RWTH-PHOENIX-Weather 2014 corpus. For example, the television announcers use the term “lowest values” (“Tiefstwerte”), while the DWD texts use the word “lowest temperatures” (“Tiefsttemperaturen”). Neither dropping these words nor creating additional glosses which literally correspond to these words could solve the issue. When looking at the unknown words, besides the use of different style and vocabulary we found many noun compounds and inflected words which occur in the corpus in another form. We therefore applied the morphological analyser Morphisto to split noun compounds and reduce inflected words to their base form. Table 6.9 shows the OOV statistics of both setups with regard to the RWTH-PHOENIX-Weather 2014 Corpus. The table shows that the number of unknown words could be greatly decreased by the mapping of words to their base form and by splitting noun compounds. After training the inverse system on the morphologically processed data and translating the equally processed DWD data, the semi-supervised training data brought some improvements, especially with regards to TER. The results including the additional language model data and semi-supervised training can be seen in Table 6.10. The results show that the addition of the monolingual data greatly improved the translation quality both with regard to BLEU and TER. The addition of the bilingual data obtained by semi-supervised training led to further improvements in BLEU and TER.
Table 6.9: OOV rates of the unprocessed DWD data and the data processed by Morphisto with respect to the RWTH-PHOENIX-Weather 2014 Corpus

<table>
<thead>
<tr>
<th></th>
<th>unprocessed</th>
<th>Morphisto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>2100</td>
<td>1269</td>
</tr>
<tr>
<td>OOV (running)</td>
<td>7907</td>
<td>3850</td>
</tr>
<tr>
<td>OOV (vocabulary)</td>
<td>1206</td>
<td>619</td>
</tr>
</tbody>
</table>

Table 6.10: RWTH-PHOENIX-Weather 2014 results including DWD data used for language modelling and semi-supervised training

<table>
<thead>
<tr>
<th></th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>TER [%]</td>
</tr>
<tr>
<td>System trained on Phoenix</td>
<td>22.5</td>
<td>69.3</td>
</tr>
<tr>
<td>+ DWD language model</td>
<td>23.0</td>
<td>68.3</td>
</tr>
<tr>
<td>+ semi-supervised training</td>
<td>23.9</td>
<td>67.1</td>
</tr>
</tbody>
</table>

6.4 Summary and Contributions

In this chapter, we dealt with the problem of scarce training resources. By applying a morphosyntactic analysis on the spoken language side, we could bridge the morphological gap resulting from different inflections and noun compounding. The baseline method (“first minimal”) was developed by Daniel Stein. The method selecting the compound split with the most likely POS sequence was developed by the author.

Moreover, using synonym data, glosses and words which are synonyms but which do not often co-occur in the training data can be aligned. The approach was developed and implemented solely by the author.

By integrating monolingual data, which is easier to obtain than annotated bilingual data, we could further improve the translation quality. The collection of the data and the experiments were solely performed by the author.

To summarise, the described methods are suitable to improve the system performance given the scarce training resources prevalent in sign language machine translation.
When examining the annotation scheme of the RWTH-PHOENIX-Weather Corpus (see Section 5.5.3), we found that particularly mouthinges contain important information for the translation process. In this chapter, we want to integrate this information into our translation system. We follow two approaches: in the first approach, we adapt the granularity of the annotation using an automatic clustering approach; in the second approach, we use a dedicated mouthing recognition system and combine the hand-based information and the mouthing information directly in the translation system.

Although our work focuses mainly on the translation from a sign language into a spoken language, in Section 7.1 we use an active appearance model to distinguish sign variants for the translation into a sign language. As the annotation of the RWTH-PHOENIX-Weather Corpus mainly focused on the manual component of a sign, variants which differ only in the facial expression were annotated with the same gloss. Thus, an avatar system (see Section 4.3) would always produce the same facial expression for each variant. By automatically detecting variants, we can thus enhance the corpus annotated with manual glosses, which can benefit an avatar animation.

In Section 7.2, we combine the active appearance model with a viseme recognition system to detect mouthinges and use them in the decoding process directly. To do so, we enhance the translation system so that it can process pairs of manual glosses and accompanying mouthinges. We thus arrive at a system that can handle more than one modality.

7.1 Clustering of Glosses According to Facial Features

In text-to-avatar translation systems, facial expressions and mouth patterns are a vital part of a natural sign language avatar animation. However, gloss-based corpora often lack detail with respect to such non-manual features. To create a translation system which can produce facial expressions and mouthinges, a more fine-grained annotation is necessary. We apply a clustering algorithm to automatically distinguish signs with the same manual component but different facial patterns using an active appearance model. The resulting translation system can then produce such patterns when translating the written language text.

7.1.1 Sign Variants

The annotation of a sign language video corpus highly depends on the task at hand. For example, if a linguist wants to study certain linguistic patterns, the annotation should be detailed with respect to these patterns. In the same way, an annotation suitable for an automatic sign language
recognition system should be tailored according to the features which the system can actually recognize. Since the RWTH-PHONIX-Weather corpus was originally developed for the recognition of hand-based features, both the time boundaries of the ID-glosses and their label were mainly based on the signing hands. This means that signs which are identical in the hand components but differ in their mouthing received the same label. For example, in German Sign Language the sign of a specific mountain is often formed by performing the general sign for mountain with the hands and silently pronouncing the name of the mountain (see Figure 7.1). In the corpus, all sign variants are glossed as “MOUNTAIN”. Using the corpus in a text-to-avatar scenario implies that the system can only reliably produce hand patterns but not other features such as facial expressions or mouthings.

In order to create a translation system which distinguishes between signing variants featuring different mouthings and facial expressions, the gloss annotation has to be more fine grained.

An optimal solution to this granularity problem would be the manual refinement of the annotation, but this process would be both time-consuming and expensive. Moreover, as one focuses on more and more aspects of sign language, the annotators would need to refine the annotation again and again. We therefore developed an automatic clustering technique which distinguishes glosses with different facial patterns based on their translation in the spoken language.

### 7.1.2 Outline of the Algorithm

The standard approach to a gloss-based text-to-avatar system is to translate the spoken language text into a sequence of glosses and then to create an avatar animation based on the glosses. The issue is that a gloss is always animated in the same way, because the context information is lost, and gloss variants are not considered.

In our approach, when translating the spoken language text into glosses, the translation system output also retains the original spoken word which corresponds to the gloss, thus consisting of (gloss, translation) pairs. E.g., if the system translates the spoken word “Alps” into the gloss MOUNTAIN, it produces the pair (MOUNTAIN, Alps). The clustering algorithm then determines a suitable sample video with the correct mouthing and facial expression to animate each specific gloss variant, based on an active appearance model which detects these expressions. Note that not
7.1 Clustering of Glosses According to Facial Features

all signs have variants, and not all signs are accompanied by mouthings or facial expressions. In this case, the system should only generate one gloss variant. Here is an outline of the procedure:

Input: a gloss-annotated video training corpus with corresponding sentences in a spoken language

1. Each gloss is aligned to at most one corresponding word in the spoken language using the open-source toolkit GIZA++. With this alignment, each gloss is provided with its translation in the spoken language as additional context information, leading to (gloss,translation) pairs. This information is then used to guide the clustering process.

2. For each (gloss,translation) pair, all videos labelled with this pair are clustered:
   a) Facial features of all videos are extracted using active appearance models.
   b) The similarity of the facial features and mouthing between pairs of glosses is calculated based on a hidden Markov model, resulting in a distance matrix.
   c) The videos are clustered according to the distance of the facial features.
   d) The central element (medoid) of the biggest cluster is selected to obtain an appropriate facial expression and mouthing variant when translating a spoken language word into a gloss.

Output: for each (gloss,translation) pair, a representative video is provided from which facial features can be extracted for avatar animation.

The clusters obtained by the algorithm are variants of a sign exhibiting different facial patterns. For signs with no or only one mouthing, the output contains only one cluster. If a cluster obtained by this procedure consists of only one or a few glosses, it can be considered an outlier or a seldom variant. The heuristic of the algorithm is to select the biggest cluster, because it was the variant which was seen most often in the context of the original spoken word. Within the cluster, the medoid is selected as the representative video. Thus, on the one hand the algorithm distinguishes different facial features and mouthings, and on the other hand it helps avoiding less standardized variants not suitable for animation.

The above algorithm can be performed in the training phase. When translating a new text from the spoken language, the spoken words are translated into glosses, and for each translated gloss the aligned word is stored. For each (gloss,translation) pair, the algorithm has determined a unique video for the facial animation of the avatar. In comparison, a standard translation system would animate all occurrences of a gloss with the same facial features and would not be able to distinguish gloss variants such as “mountain” and “Alps”.

In the following sections, we discuss the algorithm in detail.

7.1.3 Alignment

Since the mouthings of a sign often mimic the words of the spoken language, providing the spoken word as a context can help to select a sign with a specific mouthing. We therefore align the glosses to the spoken language text in order to obtain the meaning of a gloss in a given context. We use the open-source toolkit GIZA++ to align each gloss to at most one word. This process leads to a set of (gloss,translation) pairs. For each instance of such a pair, we also have the corresponding video of the persons signing, as each gloss in the corpus has manually annotated time boundaries.

To extract the facial features and mouthing for such a pair, we need to select a representative video from this set of videos. This leads to two problems. First, one (gloss,translation) pair might have several variants with respect to facial expression and mouthing. The variants might be caused by
regional dialects or personal preferences. Second, some videos might be of a poor quality and not suitable for extracting features, e.g. if the mouth is occluded by the signing hands. To solve both problems, we cluster the videos with respect to their facial features which are extracted using an active appearance model.

### 7.1.4 Active Appearance Models

To train active appearance models (see Section 3.4 for an introduction) on the RWTH-PHOENIX-Weather Corpus, 38 facial landmarks for all seven interpreters have been labelled in a total of 369 images (that is, about 50 images per interpreter). Care was taken in selecting a set of images which contain many different expressions, including extreme ones, such that the trained models can approximately represent a large span of expressions for each interpreter. Two examples of such facial annotations can be seen in Figure 3.15 of Section 3.4.

Seven SICAAMs (simultaneous inverse-compositional active appearance models) specific to the seven interpreters of RWTH-PHOENIX-Weather have been trained for the end purpose of extracting high-level facial features from the gloss-annotated videos as shown in Figure 3.17. Training and tracking with one single SICAAM for all seven interpreters would have been a viable choice as well because of the enhanced robustness of this AAM variant to variability in identity. However, we wanted to obtain the best possible accuracy in the tracking of the low-level point features. On the other hand, the calculation of the high-level features is rather sensitive to identity changes and as such had to be designed in an identity-dependent fashion. The extraction of reliable identity-independent facial features similar to those used in this work is part of the advanced computer vision research topic known as “expression transfer” and is beyond the scope of this section, where our primary goal is to give a proof of concept that gloss-based corpora can be enhanced by automatic face analysis methods.

Using the output of the active appearance model can be used to drive the animation of an avatar’s face. The grid of fitted AAM shape points shown in the top-left part of Figure 3.17 have known positions in 3D-space, as one can see illustrated in the top-right part of the figure where the grid has been normalized in 3D to get a frontal pose. These accurate point positions along with the high-level features extracted from them (shown at the bottom of the figure) convey all the necessary information for modelling and transferring continuously facial expressions to an articulated avatar’s face, using mapping techniques such as the ones proposed in [Saragih & Lucey+ 11] where a geometrical transfer matrix (from the deformable shape model to the avatar’s control nodes) is combined with a higher-level, semantic transfer map.

### 7.1.5 Similarity Metric for Facial Expressions

Applying the active appearance models to the videos of individual signs, we extract a time series of higher level features. Note that we use the manually annotated time boundaries of the signs to obtain these videos. We then use the publicly available open source speech recognition system RASR [Rybach & Gollan+ 09] to model these sequences of higher level features. This approach allows us to automatically calculate the degree of similarity between pairs of videos, as described below. We store the distances between two videos (distances being a measure of dissimilarity) in a global distance matrix, which is then used to cluster all videos labelled with the same (gloss,translation) pair.

The distance is calculated as follows: we model each facial feature by a separate Hidden Markov model (HMM), which constitutes a stochastic finite state automaton. The number of states $S_p$ for all models belonging to a (gloss,translation) pair $P$ is chosen based on the actual length of the feature sequence $x^T$. The video with the shortest sequence of frames determines the
7.1 Clustering of Glosses According to Facial Features

Co-articulation effects are accounted for by a single state garbage model which can optionally be inserted at the beginning or end of a sequence. Single Gaussian densities, a globally pooled covariance matrix and global state transition penalties are used with maximum likelihood training [Rybach & Gollan+ 09]. This is done in a nearest neighbour fashion, such that each video is represented by one model. The free HMM parameters, such as the time distortion penalties, are optimized in an unsupervised manner using the German translation as weak labels.

7.1.6 Clustering of Sign Variants

For clustering, we apply the adaptive medoid-shift clustering algorithm [Asghar & Rao 08]. The algorithm clusters the elements around medoids, i.e. representative samples for each cluster by assigning each element to the cluster of the nearest medoid. By selecting the biggest cluster, we avoid outliers which are separated into smaller clusters. Moreover, we select the medoid of the biggest cluster to obtain a video which is representative of the whole cluster. The facial features of this video can then be extracted to improve the animation of facial expressions of an avatar system.

7.1.7 Evaluation

To evaluate the clustering algorithm on the RWTH-PHOENIX-Weather Corpus, we additionally labelled the mouthing for a subset of glosses. Starting with the most frequent glosses, we selected a subset of 23 glosses (2.5% of the vocabulary) for which more than one mouthing exists in the corpus. For these glosses, we select the pairs of glosses and their aligned spoken language word which were seen more than five times. This led to 64 (gloss,translation) pairs. For each pair, we labelled up to 25 instances in the corpus, though for many pairs the number of instances in the corpus was smaller. In total, we labelled 640 running glosses. The labels are not used in training but solely for the purpose of evaluating the quality of the resulting clusters.

On average, each (gloss,translation) pair has 2.1 variants. This includes cases where a sign is sometimes accompanied by mouthing and sometimes by a mouth gesture. In addition to evaluating the clustering, we also want to evaluate the quality of the medoid by checking whether the medoid, i.e. the representative video, has the same mouthing as the videos in the same cluster.

For each (gloss,translation) pair, we calculate the precision, recall and f-measure. In Figure 7.2, we plot the distribution of these measures by sorting the scores in ascending order. The
plots show the distribution of precision, recall and f-measure between the clusters provided by the algorithm and the hand-labelled mouthings for each (gloss,translation) pair. On average, about two thirds (65.3%) of the (gloss,translation) pairs are correctly classified.

Besides the quality of the clustering, we are mainly interested in whether the adaptive medoid-shift algorithm selects a good representative video. For this, we also labelled the medoids resulting from the above clustering. The accuracy of the selected medoids is the fraction of the labelled data which has the same label as the medoid of the cluster they are in. The distribution of the accuracy is presented in Figure 7.3. On average, the algorithm has an accuracy of 78.4%, which means that in about four of five cases, the algorithm selects a representative video with the correct facial expression or mouthing.

One issue which led to errors in our method are the manually annotated time boundaries. We found that the time boundaries were mainly annotated based on the manual signs. However, there are cases in which the mouthing occurs with an offset, which means that the manual sign has finished but the mouthing still continues. For a linguistic discussion of such “spreading behaviour” which occurs in many sign languages, see [Crasborn & van der Kooij + 08]. In such cases, the mouthing could not be recognised, because the time boundaries cut off the end of the mouthing.

Figure 7.4 shows a successful clustering. The two image sequences were extracted from the corpus. The upper sequence shows the sign “Allgäu” (a hilly region in southern Germany) in which the hands perform the sign for mountain and the word “Allgäu” is pronounced. The lower image sequence in the same figure shows the medoid of the cluster the upper video was placed into. The example shows that the algorithm is able to recognize similar mouthings between different signers even if they sign and mouth at different speeds. In the overall text-to-avatar pipeline, the word “Allgäu” would be translated into the gloss MOUNTAIN, and the suitable mouthing “Allgäu” would be selected for avatar animation.

7.1.8 Summary

In this section, we approached the issue of multimodality in sign languages by detecting sign variants with the same manual component but different facial expressions and mouthings. By clustering videos with similar facial expressions and selecting a representative video for the extraction of facial features for avatar animation, we went beyond the state of the art of gloss-based translation and avatar systems which always animate a gloss in the same way. In the field of corpus annotation, the algorithm can be used to distinguish gloss variants and create a more fine-grained
7.2 Incorporating Multimodality Into Decoding

In this section, we want to go beyond a single-modality gloss-based translation system by using the output of a sign language recognition system using hand-based features and a mouthing recognition system and combining the two information streams in the translation system.

Since the approach was generally designed for both mouthings and facial expressions, whole signs were modelled in the HMM system. For the special but important case of mouthing recognition, we can go beyond this limitation and model the mouthing as a sequence of visemes, i.e. visual phonemes. The advantage of such a model is that the number of samples per viseme is much larger than the number of videos per (gloss,translation) pair, leading to better model estimates. For the translation from a sign language into the corresponding spoken language, we use a viseme recognition system and combine the hand-based glosses with the recognised mouthing directly in the translation system.

7.2 Incorporating Multimodality Into Decoding

Since the approach was generally designed for both mouthings and facial expressions, whole signs were modelled in the HMM system. For the special but important case of mouthing recognition, we can go beyond this limitation and model the mouthing as a sequence of visemes, i.e. visual phonemes. The advantage of such a model is that the number of samples per viseme is much larger than the number of videos per (gloss,translation) pair, leading to better model estimates. For the translation from a sign language into the corresponding spoken language, we use a viseme recognition system and combine the hand-based glosses with the recognised mouthing directly in the translation system.
Sign language translation is very similar to spoken language translation as it combines a
recognition system and a translation system (see Figure 7.5 (a) and (b)). As we discussed in
Section 7.1.1, the annotation of the RWTH-PHOENIX-Weather Corpus is mainly based on the
manual signing. Thus the recognition system is mainly a “hand sign recognition system”.

In the approach proposed in this section, which is depicted in part (c) of the same figure, we
add the mouthing information obtained from the viseme recogniser (described in Section 3.5) as
an additional knowledge source to the translation system. This is done in the following way. In
cases in which the viseme recogniser has a high confidence to recognise a word correctly, we split
up the gloss into several variants. E.g. the manual gloss MOUNTAIN(="BERG") from Figure
7.1 could be split up into two gloss variants MOUNTAIN_alps and MOUNTAIN_mountain. The
machine translation system is then trained on these gloss variants.

7.2.1 Mouthing Features in Machine Translation

Since the mouthing usually corresponds to a word in the spoken language, we want to increase
the probability of the gloss variants which are translated into their mouthing component. This
can be done on the word and the phrase level, leading to additional features in the translation
system

On the word level, we increase the probability of the IBM1-like lexical smoothing of pairs
where mouthing and translation coincide by a factor $\alpha$:

$$N_s(e, f) = \sum_{e_i: e_i = e} \sum_{f_{j_s}=f} \sum_{\{a_i\}_s} \beta \frac{1}{|\{a_i\}_s|}$$

$$N(e, f) = \sum_s N_s(e, f)$$

$$p(e|f) = \frac{N(e, f)}{\sum_{e'} N(e', f)}$$

$$\beta = \begin{cases} 
1 + \alpha & \text{if } e_i = \text{mouthing}(f_{j_s}) \\
1 & \text{else}
\end{cases}$$

The factor $\alpha$ is optimized on the development set. An example sentence for lexical smoothing
can be found in Figure 7.6.

On the phrase level, we add binary as well as count features to the phrase table, indicating
whether a gloss with a certain mouthing is translated into the corresponding spoken word (boolean
feature) or counting the number of glosses in the phrase for which this is true (count feature). In
the example in Figure 7.6, a phrase consisting of the whole sentence would receive an indicator
feature of 1 and a count feature of 3, since there are three matches of mouthings and translations.

Thus, the computer would e.g. learn to translate the gloss variant MOUNTAIN_alps (which
consists of the manual sign for mountain, accompanied by the mouthing “Alps”) into the German
word for Alps. We refrained from hard-wiring these connections for two reasons. First, the viseme
recognition also contains errors, which can partly be learnt by the machine translation system
during training. Moreover, mouthings usually use the base form of the word without inflections,
and thus the same mouthing can result in different inflections in the spoken language.

7.2.2 Setup of Multimodal System

In this section, we briefly summarise the setup of the multimodal system including a manual sign
recognition system and a viseme recognition system based on a manual gloss corpus.
7.2 Incorporating Multimodality Into Decoding

1. Align glosses to at most one word in the spoken language using GIZA++ to obtain possible mouthings.

2. Obtain a pronunciation lexicon for the aligned spoken words, using a grapheme-to-phoneme conversion system.

3. Map the pronunciation lexicon to visemes to obtain a viseme lexicon.

4. Train the viseme recognition system on the glosses and the corresponding viseme pronunciation of the aligned words.

5. On the development set, measure the error rate of the viseme recognition system for each gloss and only create gloss variants for glosses with a low error. Glosses with a high error do probably not have different mouthings.

6. Train the machine translation system on the gloss variants including the feature functions described in the previous section.

7. On the test set, use the viseme recognition system to recognise the mouthings and translate the glosses or gloss mouthing variants.

Note that steps 1-5 can also be used to automatically annotate the mouthing of a corpus which has only manual glosses.

7.2.3 Experiments

First we examine translation results assuming that all mouthings have been recognized correctly. The mouthings are obtained by aligning the glosses to at most one spoken word in the reference using GIZA++. The results thus have to be considered as oracle results, because they use the reference texts, but they provide an upper bound on the translation performance of the actual system and show the potential of adding the mouthing information to the system. The results can be seen in Table 7.1. The baseline system uses only the manual glosses, but includes the techniques described in Chapter 6 such as morpho-syntactic analysis and usage of the synonym...
Table 7.1: Machine translation results based on mouthing transcriptions, assuming all mouthings were recognized correctly

<table>
<thead>
<tr>
<th>System</th>
<th>Dev BLEU</th>
<th>Dev TER</th>
<th>Test BLEU</th>
<th>Test TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual glosses</td>
<td>18.6</td>
<td>71.8</td>
<td>23.2</td>
<td>72.6</td>
</tr>
<tr>
<td>Manual glosses + GIZA-aligned mouthings</td>
<td>36.8</td>
<td>53.4</td>
<td>29.9</td>
<td>59.8</td>
</tr>
<tr>
<td>+ word level feature function</td>
<td>39.8</td>
<td>45.3</td>
<td>31.7</td>
<td>52.7</td>
</tr>
<tr>
<td>+ phrase level feature function</td>
<td>40.8</td>
<td>43.6</td>
<td>32.6</td>
<td>49.9</td>
</tr>
<tr>
<td>+ word + phrase level feature function</td>
<td>41.1</td>
<td>44.4</td>
<td>33.6</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Table 7.2: Machine translation results of systems including viseme recognition input

<table>
<thead>
<tr>
<th>System</th>
<th>Dev BLEU</th>
<th>Dev TER</th>
<th>Test BLEU</th>
<th>Test TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual glosses</td>
<td>18.6</td>
<td>71.8</td>
<td>23.2</td>
<td>72.6</td>
</tr>
<tr>
<td>Manual glosses + Viseme Recognition</td>
<td>35.2</td>
<td>53.2</td>
<td>23.1</td>
<td>73.4</td>
</tr>
<tr>
<td>+ word level scaling factor</td>
<td>36.1</td>
<td>54.3</td>
<td>24.1</td>
<td>70.9</td>
</tr>
<tr>
<td>+ phrase level scaling factor</td>
<td>36.8</td>
<td>53.5</td>
<td>24.4</td>
<td>69.6</td>
</tr>
<tr>
<td>+ word + phrase level scaling factor</td>
<td>37.5</td>
<td>52.6</td>
<td>24.8</td>
<td>69.3</td>
</tr>
</tbody>
</table>

database to improve the corpus alignment. It is therefore better than the result described in Table 5.7. Training a phrase-based system on the gloss-variants increases the system performance by 6.7 BLEU and 12.8 TER. This indicates that a system based only on manual glosses misses a lot of important information, as we already discussed in Section 5.5.3. Additional gains can be obtained by increasing the probabilities of matching mouthings and translations on the word and phrase level. The best performance can be obtained by combining both of these models.

The translation result of the whole pipeline of viseme recognition and translation system is given in Table 7.2. Training the machine translation system on the gloss variants produced by the viseme recognizer first led to a degradation due to the errors in the viseme recognition. However, increasing the weight of corresponding mouthing and translation pairs either on the word or the phrase level led to an overall improvement over the baseline. Combining both models further led to slight improvements both BLEU and TER further.

Looking at the translation results and comparing the system which uses only the manual glosses to the best system which uses both these glosses and the viseme recognition output, we found that the system mainly improved sentences containing at least one of a small set of glosses which are quite common in weather forecasting and which have variants similar, but distinct in meaning, e.g. COLD (mouthing: cold) vs. COLD (mouthing: frost), or EVENING (mouthing:evening) vs. EVENING (mouthing:night). While the hand-based gloss system cannot distinguish these variants and thus produces one or the other translation with similar probabilities,
Table 7.3: Example sentence of the translation systems using (a) only the manual glosses and (b) the manual glosses and the viseme recognition. Below the middle bar, there are English translations for reference. The system using the viseme recognition output can distinguish the sign variants “cold” and “frost”.

<table>
<thead>
<tr>
<th>gloss</th>
<th>mouthing</th>
<th>PLUS 6</th>
<th>TEMPERATURE VOR POMMERN LOC</th>
<th>BISSCHEN</th>
<th>KALT</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>am vormittag</td>
<td>bis</td>
<td>plus 6 grad vorpommern</td>
<td>ein bisschen</td>
<td>kuehler</td>
</tr>
<tr>
<td></td>
<td>bis plus 6 grad</td>
<td>vorpommern</td>
<td>unter leichtem</td>
<td>frost</td>
<td></td>
</tr>
<tr>
<td>reference</td>
<td>Dort morgen bis plus 6 Grad, in Vorpommern gibt’s leicht</td>
<td>frost</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>gloss</th>
<th>mouthing</th>
<th>PLUS 6</th>
<th>TEMPERATURE WESTERN POMERANIA LOC</th>
<th>LITTLE</th>
<th>COLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>in the morning</td>
<td>up to plus 6 degrees western pomerania</td>
<td>a bit colder</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>under light</td>
<td>frost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reference</td>
<td>There tomorrow up to plus 6 degrees, in Western Pomerania</td>
<td>there is light frost</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the combined system can distinguish the variants. An example can be seen in Table 7.3. The mouthing recognition correctly detects the mouthing “frost” for the manual sign COLD and thus can produce a better translation.

7.2.4 Summary

In this section, we approached the issue of multimodality by combining hand-based glosses and the output of mouthing recognition system and using both information as an input to a statistical machine translation system. While the mouthing recognition system is still far from perfect, the combination of the input could improve the translation quality compared to the translation of only hand-based glosses.

7.3 Corpus-NGT

While the main focus of our work in the SignSpeak project was on the RWTH-PHOENIX-Weather Corpus, we also spent some time working on the Corpus-NGT provided by our project partner Radboud University Nijmegen. The corpus is interesting for several reasons: first, the domain is broader, but also more challenging, consisting of conversations about Deaf-related issues. Moreover, the annotation consists of glosses, but each hand is annotated on an individual tier (see Figure 5.1), highlighting the parallel nature of sign languages. With this annotation, it is known whether the signer used both hands for a sign or which hand was used for one-handed signing. In the following sections, we examine whether we can exploit this information in machine translation. In Section 7.3.3, we also examine the effect of head shakes on the translation quality.

7.3.1 Corpus Description

[Crasborn & Zwitserlood 08] presents a data collection in the Sign Language of the Netherlands. It consists of recordings in the domain of fables, cartoon/home video paraphrases, discussions on various topics and free conversation. After a careful examination of the data, we excluded the topics “funniest home videos”, fables and “Tweety & Sylvester” because of their huge amount of
productive signing,\textsuperscript{1} which poses a daunting task to both automatic sign recognition and translation. Further, we discarded free conversations as well as talks about personal experience because they had an average type/token ratio of only 3.2 and 4.8, respectively. In contrast, the domain of discussions on selected topics that are related to deafness and Deaf culture had an average type/token ratio of 8.5 and 6.0, since the vocabulary was somewhat restricted due to the specific questions the signers argue about. In this setting, two signers are sitting face to face and discuss a topic that was shown to them in form of a written question. The translations of the sentences into spoken Dutch were provided by the Radboud University, Nijmegen, as part of the EU-project SIGNSPEAK.

Table 7.4: Corpus statistics for Corpus-NGT

<table>
<thead>
<tr>
<th></th>
<th>Right Hand</th>
<th>Left Hand</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>1699</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Running Words</td>
<td>8129</td>
<td>4123</td>
<td>15130</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1066</td>
<td>773</td>
<td>1695</td>
</tr>
<tr>
<td>Singletons</td>
<td>481</td>
<td>376</td>
<td>840</td>
</tr>
<tr>
<td>Test:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentences</td>
<td>175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Running Words</td>
<td>875</td>
<td>496</td>
<td>1815</td>
</tr>
<tr>
<td>Distinct Words</td>
<td>272</td>
<td>181</td>
<td>426</td>
</tr>
<tr>
<td>OOVs</td>
<td>46</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Trigram perpl.</td>
<td>107.0</td>
<td>54.6</td>
<td>67.5</td>
</tr>
</tbody>
</table>

The Corpus-NGT can be considered far more challenging than the RWTH-PHOENIX-Weather 2014 corpus. With the status of the corpus at the end of the project (see Table 7.4 for the corpus statistics excluding fables and free conversations), we do not expect to reach satisfying translation error scores. This is mainly due to the much broader domain when compared to weather forecasting, and due to the conversational, even casual nature of the signed sentences. Hesitations and partial sentences are frequent, and some information is only conveyed by non-signed (and thus non-glossed) communication channels such as facial expressions (e.g. “I totally agree”), which is still translated into Dutch. However, the corpus is interesting from a scientific point of view for the following reasons. First, we can assume the grammar of the sign language to be more accurate, since it is derived from close-to-natural conversations among Deaf. Apart from the written questions that start the discussions, the Dutch grammar does not have an immediate impact on the word order or the communication devices used. Second, the discussion domain is much broader than the weather forecasting domain, which is more interesting for statistical machine translation, since limited-domain problems could in general also be solved by rule-based systems, which is however not feasible for larger domain tasks as the rule set to be created would become too large. Third, the annotation procedure used in this corpus better reflects the parallel nature of sign languages. In this corpus, each hand is annotated individually on a different tier (see Figure 7.7), and time-aligned head shake annotations were included for a subset of the corpus. The gloss annotation was not influenced by the text of the spoken language, as the translations into spoken Dutch were only added later.

\textsuperscript{1}Productive signing refers to the creation of new signs to describe shape, location, movement or other aspects of objects or persons. Closely related to this is the concept of classifiers described in Section 3.2.2.
7.3.2 Parallel Input Channel Approaches

In [Stein & Schmidt 12], we analysed the question of how to handle input where the two hands are glossed individually. The example in Figure 7.7a shows that for some sentences, the dominant hand covers all words of the sentence and the non-dominant hand remains motionless for signs that only require one active hand. However, this is not always the case. The example in Figure 7.7b shows the transcription of a signer who switches the active signing hand within one sentence.

(a) “It is quite hard to communicate with hearing persons.”

(b) “If you haven’t been to the club house for some time, you will not know the sign for the word ‘email’.”

Figure 7.7: Example sentences from the Corpus-NGT corpus. Each hand is annotated on a separate tier.

We performed three experiments. First, we only employed the right hand information as our source input data and define this as our baseline. A next approximation is to select for each sentence the glosses of the hand that signs more words, an approach which we call active hand. In a third step, we parsed the annotation file again, matched the timing of the individual glosses, and time-aligned both gloss tiers, omitting word duplications whenever both hands sign the same (merged hands). Note that this method still does not capture the whole expressiveness of sign languages, e.g. a signer might sign “NEWSPAPER” with both hands, keeping the non-dominant hand in this position but signing “COFFEE” with his dominant hand in the meantime, a signed construction which could be translated as “drinking coffee while reading the newspaper”.

The results can be found in Table 7.5. For the phrase-based system, switching from the right hand to the active hand gives an improvement of 1.4 in BLEU, and the merged hand approach further improves the BLEU score by 1.9. Both improvements are statistically significant (p < 0.01). For the hierarchical system, the improvements over the baseline were smaller.

In this section, we further examine the Corpus-NGT with regard to the issue of multiple input streams. To assess whether the merging of the input streams is a suitable option or whether a modelling of independent streams is a necessity, we calculated statistics for the temporal co-occurrence of glosses for the two hands. Figure 7.8 shows the different possible types of temporal co-occurrence. Either only the left or the right hand signs, or both hands sign at the same time. Simultaneous signing can be overlapping, where one hand starts and ends earlier than the other, or including, when one hand begins after and ends before the other hand. We further distinguish the cases of both hands signing the same sign or different signs.
Table 7.5: Corpus-NGT translation results applying different strategies for using two gloss streams

<table>
<thead>
<tr>
<th></th>
<th>Phrase-Based</th>
<th>Hierarchical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU [%]</td>
<td>TER [%]</td>
</tr>
<tr>
<td>Right Hand</td>
<td>8.0</td>
<td>77.5</td>
</tr>
<tr>
<td>Active Hand</td>
<td>9.4</td>
<td>76.1</td>
</tr>
<tr>
<td>Merged Hands</td>
<td>11.3</td>
<td>76.8</td>
</tr>
</tbody>
</table>

| Left hand            | right only   | synchronous | overlap   | inclusion   |
|                      |              |             |           |             |
| Right hand           | right only   | synchronous | overlap   | inclusion   |

Figure 7.8: Temporal arrangement of glosses on two input streams for the Corpus-NGT

The statistics can be found in Table 7.6. The types in the upper block can be easily modelled by merging both gloss streams, since either only one hand signs, or both hands sign the same sign, either synchronously or with some temporal variability. The types in the lower block are more problematic to be handled by the merging approach. Since both hands sign different signs at the same time, by merging both glosses into one stream the information of the temporal overlap is lost. Of the three types of temporal overlap, the inclusion is most common, occurring in 8.7% of the glosses. This means that one hand performs a sign and holds this sign while performing another sign with the other hand. However, when looking at some videos and annotations of these cases, we found that in many cases, the holding of one sign did not bear a special meaning which would be lost by merging both gloss streams. While certainly more correct from a linguistic point of view, we did not find many examples in which a dedicated parallel input would solve translation problems which are otherwise unsolvable. One reason is that we selected subcorpora of the Corpus-NGT which contain relatively few classifier constructions. And it is these classifier constructions that make most use of independent signing of both hands, because the two hands are often used to describe relations, and one hand would mark an object, whereas the other hand would describe the relation of another object. For example, to describe the geographical relation between Paris and Berlin, the left hand would mark the position of Paris on a virtual map in the signing space, while the right hand would move to the upper right to indicate Berlin lies to the northeast of Paris. However, such cases do not occur often in the Corpus-NGT because of the domain of the corpus.

### 7.3.3 Head Shakes

For a subset of the Corpus-NGT, there are not only glosses annotated independently for each hand, but also head shake annotation on an additional tier. In this section, we examine the use of head shakes as additional information for a translation system. Note that the experimental results are based on a subset and are therefore not comparable with the previous results.

The Corpus-NGT contains several different symbols for head shakes. A summary can be found in Table 7.7. The annotation distinguishes between head shakes ("N") and head sways ("Ns"). Moreover, a distinction is made whether a head shake or a head sway indicates a negation or not. For example, a person might “shake his head” to indicate that he looks to the left and the right before crossing a road, or following a tennis match. Both scenarios do not indicate any negation.
Table 7.6: Statistics for the temporal co-occurrence of signs on two hand tiers (=, ≠: left and right
glosses are identical/not identical)

<table>
<thead>
<tr>
<th>type of co-occurrence</th>
<th>counts</th>
<th>relative frequency</th>
<th>ok to merge</th>
<th>Σ rel.freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>only right</td>
<td>6601</td>
<td>53.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>only left</td>
<td>1151</td>
<td>9.4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>synchronous =</td>
<td>2075</td>
<td>16.9%</td>
<td>yes</td>
<td>88.2%</td>
</tr>
<tr>
<td>overlap =</td>
<td>316</td>
<td>2.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>inclusion =</td>
<td>680</td>
<td>5.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>synchronous ≠</td>
<td>29</td>
<td>0.2%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>overlap ≠</td>
<td>352</td>
<td>2.9%</td>
<td>no</td>
<td>11.8%</td>
</tr>
<tr>
<td>inclusion ≠</td>
<td>1061</td>
<td>8.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(annotated by symbol “Nn”). Lastly, the annotation distinguishes head shakes co-occurring with
manual signs, those which fall in between manual signs, and those with no accompanying manual
sign. The latter mostly occurs when one person is signing and the other person shakes their head
to indicate disapproval. As in the previous section, we included head shakes by merging them
with the two tiers of the manual glosses by sorting them according to the timeline. This led
to several problems, for example in the case when a person shakes his head continuously during
several consecutive sentences. In this case, head shakes were added to each sentence.

Table 7.7: Different annotation symbols for head shakes in Corpus-NGT

<table>
<thead>
<tr>
<th>Code</th>
<th>Negation</th>
<th>Description</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>yes</td>
<td>cooc. with manual signs</td>
<td>690</td>
</tr>
<tr>
<td>Nx</td>
<td>yes</td>
<td>falling in between two manual signs</td>
<td>116</td>
</tr>
<tr>
<td>Nf</td>
<td>yes</td>
<td>not cooc. with manual signs (i.e. especially of the addressee)</td>
<td>287</td>
</tr>
<tr>
<td>Nn</td>
<td>no</td>
<td>cooc. with manual signs</td>
<td>22</td>
</tr>
<tr>
<td>Ns</td>
<td>yes</td>
<td>Head sway from side to side</td>
<td>35</td>
</tr>
<tr>
<td>Nx</td>
<td>no</td>
<td>Head sway from side to side</td>
<td>22</td>
</tr>
</tbody>
</table>

We compared four setups: the first contains no head shake annotation, establishing the baseline. The second system contains all the different kinds of head shake symbols. In the third system, only the most common symbol “N” was included, whereas in the last system, all head shakes indicating negation were mapped onto one symbol, while head shakes not indicating negation were not included. The results in Table 7.8 show some improvement of the last setup with respect to BLEU and TER, which means that the distinction of different kinds of head shakes is not necessary for machine translation. This is of practical importance, as the head shakes could be automatically detected using the active appearance model described in Section 3.4, and the distinction of different head shake types is not necessary.
Table 7.8: Results of translation systems including different variants of head shake annotation

<table>
<thead>
<tr>
<th>Annotation Type</th>
<th>BLEU [%]</th>
<th>TER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No head shake</td>
<td>8.9</td>
<td>70.9</td>
</tr>
<tr>
<td>All head shakes</td>
<td>8.1</td>
<td>78.7</td>
</tr>
<tr>
<td>Only “N”</td>
<td>9.2</td>
<td>71.8</td>
</tr>
<tr>
<td>Map negations to one symbol</td>
<td>9.7</td>
<td>70.2</td>
</tr>
</tbody>
</table>

### 7.3.4 Summary

In this section, we investigated the Corpus-NGT which features discussions on Deaf related issues. We examined its use of parallel input streams for each individual hand and found that for statistical machine translation it is sufficient to merge the information for individual hands along the timeline. Only a few cases existed in the corpus in which the interaction between the hands played an important role. In such cases, a knowledge-based approach would be necessary to derive the meaning from the interaction. We further investigated the inclusion of head shakes and found that a simple annotation of one type of head shakes is sufficient for statistical machine translation. While the overall translation results were not satisfactory due to the broad domain and relatively small corpus size, the corpus is interesting to investigate due to its detailed annotation scheme.

### 7.4 Summary and Contributions

In this chapter, we approached the issue of multimodality in sign language machine translation from several angles.

First, we automatically clustered sign variants which are identical in the manual component but different in the facial expression and mouthing using an active appearance model. The clustering can be used to improve the facial animation of an avatar in a text-to-avatar translation system. Moreover, the algorithm also provides an unsupervised way to make the gloss annotation more fine-grained by detecting variants. The idea of clustering gloss variants and of aligning glosses and spoken words to obtain possible mouthings was developed by the author. The low-level features generated by an active appearance model were provided by Thomas Hoyoux. The implementation of the clustering and the evaluation on the test set were joint work with Oscar Koller.

Second, we used a viseme recognition system to recognise mouthings and combine the hand-based glosses with the mouthing information in the translation process. This combination led to improvements of +1.6 BLEU and -3.3 TER over a manual gloss based system. The experimental results show the importance of mouthing for the translation of Sign Language, especially for the case of German Sign Language. For the low level features, we again used the system by Thomas Hoyoux. The viseme recognizer was trained by Oscar Koller. The author developed the idea, created the viseme lexicon and implemented the combination of manual glosses and mouthing in the machine translation decoder and carried out the experiments. Note that the system can also be used to automatically annotate mouthings of a corpus which is only annotated with manual glosses.

Third, for the Corpus-NGT we examined the annotation of individual glosses for each hand. We found that this annotation scheme is not exploitable by statistical machine translation and that the simple approach to merge the gloss streams according to the timeline preserves the information most of the time. Complex scenarios where left and right hand represent different objects which interact with each other cannot be captured with the direct approach taken in this work, as they require more complex knowledge-based methods. We found however that the
addition of head shakes led to some improvements. The experiments on the head shakes and the statistics of temporal co-occurrence were carried out by the author of this thesis.

To summarise, the chapter is a first step to handle more than one modality in sign language machine translation, either by adapting the granularity of the annotation to consider sign variants, by combining several modalities in the translation process or by merging multiple input streams according to the timeline.
8.1 Scientific Achievements

In this section, we revisit the scientific goals defined at the beginning of this work in Chapter 2 and examine in how far we have accomplished them.

Related to sign language annotation:

- We analysed the annotation schemata of the RWTH-PHOENIX-Weather Corpus (Chapter 5) and the Corpus-NGT (Chapter 7) and found that especially the mouthing information is important in addition to the manual information. On the contrary, time boundaries for individual hands do not lead to further information for machine translation.

- We examined the effect of recognition errors on the translation quality by translating recognition output with varying error rates as well as the annotated glosses (Chapter 5).

- We conducted a human evaluation on the results of the full pipeline of a sign language recognition system and a sign language translation system and found that the errors of the recognition system strongly effect the meaning of the output, whereas errors in the translation system mostly effect the grammaticality and fluency of the resulting text (Chapter 5).

- We efficiently annotated the spoken language side of the RWTH-PHOENIX-Weather Corpus using a speech recognition system, a punctuation and a case recovery system (Chapter 5). Our approach can drastically increase the speed of annotation of a new corpus given a good speech recognition quality in the domain at hand.

Related to data scarcity (Chapter 6):

- We dealt with the problem of scarce training resources by employing monolingual data using semi-supervised training.

- We performed a morphosyntactic analysis to bridge the gap in morphology between the glosses, which are usually in the base form, and the spoken words, which are highly inflected in German.

- We further used a synonym database to improve the alignment between glosses and spoken words and phrases by detecting synonymity between them.
Related to multimodality (Chapter 7):

- Using an automatic clustering technique, we could detect suitable facial expressions and mouthings. The technique can also be used to change the granularity of the annotation itself by splitting gloss variants.

- Using a viseme recognition system in addition to the manual glosses, we handled more than one input in the translation system itself.

- By merging multiple modality streams in the case of the Corpus-NGT, we could handle head shakes and independent streams for left and right hand.

### 8.2 Publications and Joint Work

In the course of this thesis, the following scientific publications have been successfully submitted to peer-reviewed conferences and journals.

**Sign Language Corpora:**

- [Forster & Schmidt + 14] Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather

- [Forster & Schmidt + 12] RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus

In the papers, the author performed the sign language machine translation results and performed the annotation of the spoken language side, whereas the other authors worked on sign language recognition and tracking.

**Sign Language Machine Translation:**

- [Stein & Schmidt + 12] Analysis, preparation, and optimisation of statistical sign language machine translation

- [Stein & Schmidt + 10] Sign Language Machine Translation Overkill

- [Schmidt & Koller + 13b] Using Viseme Recognition to Improve a Sign Language Translation System

- [Schmidt & Koller + 13a] Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models

In the first two papers, the author performed the experiments using the phrase-based approach, while Daniel Stein performed the experiments using the hierarchical phrase-based approach. In the remaining papers, Thomas Hoyoux provided the active appearance models, Oscar Koller worked on the viseme recognition, while the author developed the idea, implemented and trained the translation features and performed the machine translation experiments. The clustering was joint work of Oscar Koller and the author.

**International Evaluation Campaigns:**

8.3 Conclusions

In the evaluation campaigns, the authors usually focus on different language pairs and systems. In the NTCIR Patent MT campaigns, the author led the work on the Japanese-English language pair. The systems were second in NTCIR-9 and first in NTCIR-10 in terms of BLEU and NIST. In IWSLT ’11, the author contributed to the French-English system. In WMT ’11, the author contributed to the English⇔German systems, and in WMT ’13 to the English⇒German system.

In this thesis, we have addressed three issues with regard to sign language machine translation. First, we have investigated a suitable annotation scheme to capture the multimodal nature of sign languages. Examining the RWTH-PHOENIX-Weather Corpus, we found that for German Sign Language, besides the manual signing mouthings play a crucial role.

Second, we handled the problem of scarce training materials by including monolingual data using semi-supervised training techniques and by improving the alignment between glosses and corresponding words and phrases using a morpho-syntactic and a semantic analysis. Here, we found that both the morpho-syntactic analysis and the use of synonym data to generate better gloss-word alignments led to strong improvements. Using additional in-domain monolingual data, which is often easier to obtain than annotated bilingual data, mainly improved the system when we trained a larger language model. Using the data to create additional bilingual data by semi-supervised training brought further improvements.

Third, to handle multimodality in sign language machine translation, we proposed three approaches: we used an automatic clustering technique to change the granularity of the annotation by distinguishing sign variants, and we combined manual signs and mouthing information directly in the translation system. For the Corpus-NGT, which features a very detailed annotation including glosses for each individual hand and for head shakes, we found that merging this information along the time line is sufficient for statistical machine translation.

The improvements of scarce data handling and the inclusion of more than the manual component led to a translation system that is well beyond the state of the art in gloss-based sign language machine translation.
List of Figures

1.1 Vauquois triangle ......................................................... 1
1.2 Architecture of the sign language recognition and translation system developed in the SignSpeak project ......................................................... 3

3.1 Alignment between a German sentence and its English translation. The black squares indicate correspondences between German and English words. ............... 9
3.2 System architecture using a log-linear model combination. The system is easily extensible by adding additional feature functions. ................................. 11
3.3 Alignment between a German sentence and its English translation. The blue dashed rectangles indicate some valid phrase pairs. Note that valid phrases can overlap or contain other valid phrases. .................................................. 12
3.4 Separable verbs in German are one example of brace structures which cannot be learned in the phrase-based approach. .................................................. 14
3.5 Hierarchical phrases can capture different sentence structures. In German relative clauses, the verb is in the last position, while English subordinate clauses follow the regular subject-verb-object structure. .................................................. 14
3.6 Illustration of the cube pruning algorithm .................................................. 16
3.7 Graphical representation of the applied optimisation methods. The standard method withholds a part of the training data as development set. The improved method combines several systems, each withholding a different development set, similar to the method of n-fold cross-validation. .................................................. 19
3.8 Facial expressions can express degrees/amounts. The examples are in DGS, figures taken with permission from www.visuelles-denken.de .................................... 21
3.9 An example of two signs in German Sign Language differing only in their handshape. Left meaning: “to say” (“1-handshape”), right meaning: “to ask” (“F-handshape”) (taken from [Braem 95], p.19) .................................................. 22
3.10 The ASL sign “line” as an illustration and in Stokoe notation. (Dictionary entry from [Byrom 08], p. 60) The notation $\text{B}_a G_x^\perp$ indicates: the left hand has a flat hand shape ($\text{B}$) facing up ($a$) and is under (\(\perp\)) the right hand, which forms a pointing finger ($G$) touching ($x$) the left palm and moving away ($\perp$) from the signer. ............... 23
3.11 The sign “to say” (German: “sagen”, see Figure 3.9 left) in HamNoSys notation. Symbols from left to right: The handshape depicts a hand with pointing index finger (left), which can alternatively be bent (right). The orientation indicates that the hand is pointing upwards and the palm is facing diagonally to the right of the signer. The location is at the chin (left), and in contact (right) with the chin. The action indicates a movement forward and downward (1. symbol) in a circular movement (2. symbol) which is due to an elbow rotation. After (3. symbol) the downward movement, the index finger is pointing diagonally to the left (4. symbol).

3.12 The sign “question” (German: “Frage”) in SignWriting notation. The four circles indicate a temporal sequence of the mouthing of the word “f-r-a-g-e”. The handshape is iconic and similar to the drawing in Figure 3.12. The arrow indicates a movement horizontally away from the mouth, the horizontal direction is encoded in the colour of the arrowhead and its stroke. The initial contact of the hand at the lips is indicated by the asterisk.

3.13 Architecture of a continuous sign language recognition system.

3.14 Hidden Markov model with Bakis topology, to model signs performed at different speed. The depicted probabilities are state transition probabilities.

3.15 Visualisation of facial annotations.

3.16 Active appearance models are used to track landmarks on the face (green grid lines) in a video sequence. The coloured bars indicate the degree of mouth openness (red; third bar), eye openness (blue; second, fourth bar) and eyebrow raise (yellow; first, fifth bar).

3.17 High-level feature extraction: features are extracted by tracking the face in the video (a), rotating the grid into frontal position (b), and measuring distances between the points (c).

4.1 System architecture of a system translating (a) from a sign language and (b) into a sign language.

4.2 Example of a sign language avatar system: PAULA ([Davidson 06]).

5.1 Corpus annotation in ELAN of the RWTH-PHOENIX-Weather and the Corpus-NGT.

5.2 Screenshot from the RWTH-PHOENIX-Weather Corpus featuring weather forecasts interpreted into German Sign Language.

5.3 An example of a mismatch between the spoken announcement and the sign language interpretation in the RWTH-PHOENIX-Weather Corpus is shown in (a). The sign for “southwest region” performed by the interpreter instead of the region name “Breisgau” can be seen in (b).

5.4 Semi-automatic processing pipeline to obtain spoken language side of the RWTH-PHOENIX-Weather 2014 Corpus.

5.5 Example of the alignment crunching technique, taken from the RWTH-PHOENIX-Weather 2014 corpus, on the sentence “Tagsüber bleibt die Wolkendecke in Süddeutschland geschlossen.” (Engl.: “During the course of the day, the cloud cover in southern Germany remains dense.”). The word “Wolkendecke” (“cloud cover”) is a singleton, but “Wolke” (“cloud”) is of course well-known and “Decke” is known from “Schneedecke” (“snow cover”). Thus, in (a) the alignment has errors, but for compound split German in (b) the quality is much better. After crunching the alignment in (c), the alignment structure matches the original German sentence.
5.6 Translation results on the output of the recognition system of the RWTH-PHONIX-Weather 2012 corpus. A recognition error rate of 0% corresponds to the translation of the transcribed glosses, the maximum error rate of 54.4% to the translation of the actual recognition system output. Intermediate results are simulated by reducing the perplexity of the recognition system’s language model.

5.7 Human Evaluation of the translation results of the RWTH-PHONIX-Weather 2012 Corpus: adequacy and fluency scores are indicated by the bars, the vertical lines indicate the variance between the evaluators.

5.8 Correlation between the automatic metric TER and human evaluation scores.

6.1 Example of an improved alignment by using a synonym database. The left sentence has quite a poor alignment quality. On the right, the alignments of STELLENWEISE to mitunter and of VERSCHWINDEN to lichten are recognised, because the words are found in the synonym data base.

6.2 Graphical representation of semi-supervised training. The monolingual text data is translated using an inverse machine translation system (dashed arrow). The monolingual data and its translation are then used as additional training data for a new machine translation system.

7.1 Two signs with the same manual component, differing only in the mouthing. The signs are annotated with the same gloss MOUNTAIN. At the time of the snapshots, the underlined letters are pronounced.

7.2 External evaluation of the clustering with respect to the labelled data.

7.3 Accuracy of the selected medoid with respect to the labelled data.

7.4 Key frames of a sign labelled with the mouthing “Allgäu” and the medoid of the corresponding biggest cluster, slightly trimmed to focus on the mouthing.

7.5 System architectures of (a) spoken language translation, (b) (manual) sign language translation and (c) the system proposed in this section combining a manual sign recognition and a mouthing recognition.

7.6 Example of the lexicalised smoothing. The alignments with a dashed blue frame receive a higher lexical smoothing, because mouthing and translation are identical. (Right figure: English translation for reference).

7.7 Example sentences from the Corpus-NGT corpus. Each hand is annotated on a separate tier.

7.8 Temporal arrangement of glosses on two input streams for the Corpus-NGT.
List of Tables

3.1 Sign language articulators which can be used simultaneously ........................................ 20
3.2 High-level facial features used for clustering and viseme recognition and the underlying lower-level point features (see Figure 3.15) .................................................. 28
3.3 Phoneme-viseme mapping. Phonemes in one row cannot be visually distinguished. 33
3.4 Viseme Error Rate (VER) and recall of viseme recogniser measured on 640 manual annotations. .......................................................... 34

5.1 Annotation Scheme of the RWTH-PHOENIX-Weather 2014 Corpus (glosses translated into English) .............................................................. 45
5.2 Example for the application of preprocessing rules (Upper group of rows: original German glosses, lower group of rows: English translation for reference) ..... 46
5.3 Statistics of the RWTH-PHOENIX-Weather 2014 Corpus ........................................ 47
5.4 Errors of the automatic processing pipeline .............................................................. 50
5.5 The effect of the sign language recognition quality on the machine translation results, measured on the RWTH-PHOENIX-Weather 2012. ......................... 53
5.6 Translation results on the RWTH-PHOENIX-Weather 2014 multisigner test set using different approaches and methods. .............................................. 56
5.7 Translation quality of RWTH-PHOENIX-Weather 2014 depending on the use of additional information .............................................................. 56

6.1 Morphisto split of the compound word “Grenzwetterlage” (“marginal weather conditions”) with English translation for reference ........................................ 60
6.2 An example of a morphosyntactic analysis of the word “Grenzwetterlage” (“border weather situation”). The original heuristic chooses the first analysis with the shortest length (underlined, “bordering ranting situation”). The approach proposed here chooses the analysis with the most probable parts-of-speech sequence (bold, correct split). The word “Wetter” can also mean “a betting person”. The analysis represents this as “wetten er” (V NN SUFF), a verb with a noun suffix. ........... 61
6.3 Ambiguity of the Morphisto analysis of the RWTH-PHOENIX-Weather 2014 Corpus vocabulary consisting of 2 860 words ................................................. 61
6.4 Experimental results of the improved Morphisto analysis ........................................ 62
6.5 Example sentence from the RWTH-PHOENIX-Weather 2014 Corpus. The words “Westen” (west) and “Wetter” (weather) are correctly analysed by the maximum likely POS sequence approach, but incorrectly by the first minimal heuristic (as “wesen”, to rot, and “wettern”, to scold). ................................. 62
6.6 Experimental results including synonym data ................................. 63
6.7 Left: examples of glosses and their corresponding words in the German language text. The gloss-word pairs are synonyms. Right: English translation of left table for reference. ............................................................... 64
6.8 Corpus statistics for German Meteorological Service (“Deutscher Wetterdienst”) newsletter ............................................................... 64
6.9 OOV rates of the unprocessed DWD data and the data processed by Morphisto with respect to the RWTH-PHOENIX-Weather 2014 Corpus ......................... 67
6.10 RWTH-PHOENIX-Weather 2014 results including DWD data used for language modelling and semi-supervised training ................................. 67
7.1 Machine translation results based on mouthing transcriptions, assuming all mouthings were recognized correctly ........................................... 78
7.2 Machine translation results of systems including viseme recognition input .... 78
7.3 Example sentence of the translation systems using (a) only the manual glosses and (b) the manual glosses and the viseme recognition. Below the middle bar, there are English translations for reference. The system using the viseme recognition output can distinguish the sign variants “cold” and “frost”. ................................. 79
7.4 Corpus statistics for Corpus-NGT .................................................. 80
7.5 Corpus-NGT translation results applying different strategies for using two gloss streams .......................................................... 82
7.6 Statistics for the temporal co-occurrence of signs on two hand tiers (=, ≠: left and right glosses are identical/not identical) ..................................... 83
7.7 Different annotation symbols for head shakes in Corpus-NGT .................. 83
7.8 Results of translation systems including different variants of head shake annotation 84


[Koller & Ney + 13] O. Koller, H. Ney, R. Bowden: May the force be with you: force-aligned SignWriting for automatic subunit annotation of corpora. In IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–6, Shanghai, China, April 2013.


