RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus

Jens Forster[1], Christoph Schmidt[1], Thomas Hoyoux[2], Oscar Koller[1], Uwe Zelle[1], Justus Piater[2], and Hermann Ney[1]

[1]Human Language Technology and Pattern Recognition
RWTH Aachen University, Germany
lastname@cs.rwth-aachen.de

[2]Intelligent and Interactive Systems
University of Innsbruck, Austria
firstname.lastname@uibk.ac.at

Abstract

This paper introduces the RWTH-PHOENIX-Weather corpus, a video-based, large vocabulary corpus of German Sign Language suitable for statistical sign language recognition and translation. In contrast to most available sign language data collections, the RWTH-PHOENIX-Weather corpus has not been recorded for linguistic research but for the use in statistical pattern recognition. The corpus contains weather forecasts recorded from German public TV which are manually annotated using glosses distinguishing sign variants, and time boundaries have been marked on the sentence and the gloss level. Further, the spoken German weather forecast has been transcribed in a semi-automatic fashion using a state-of-the-art automatic speech recognition system. Moreover, an additional translation of the glosses into spoken German has been created to capture allowable translation variability. In addition to the corpus, experimental baseline results for hand and head tracking, statistical sign language recognition and translation are presented.

Keywords: Sign Language, Corpus , Automatic Recognition and Translation

1. Introduction

Sign languages are the native languages of the deaf and partly of the hard-of-hearing communities world wide. Although more than 100 000 people use sign language in Germany alone, the development of assistive technologies such as automatic recognition and translation of sign language lags behind similar technologies for spoken languages. State of the art speech recognition and translation systems are based on statistical models. Large amounts of labelled data are required to learn such statistical models robustly. For sign languages, there are hardly any large data collections. Most existing data sets have been recorded for linguistic research, which usually does not require a certain phenomenon to appear many times to be studied. In most cases, linguistic data collections are not limited to a specific domain, are lab recorded, focus on a special aspect of sign language and often feature challenging video recording conditions such as strongly varying illuminations and cluttered backgrounds. Typically, this kind of data differs greatly from the language encountered outside the research lab. On the contrary, some public TV broadcasting stations such as the BBC in the United Kingdom and Phoenix in Germany feature the interpretation of parts of their programmes into sign language, which may serve as a source of data. The TV station Phoenix regularly broadcasts the major public news programmes with an additional interpretation into German Sign Language using an overlay window which shows the interpreter.

Since 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 to cover this broad domain sufficiently. Consequently, we restricted ourselves to the topic of weather forecasting. By extracting the weather forecasts from the news programme, a sufficient number of shows could be obtained. Moreover, the domain of weather forecasting features a limited vocabulary and a restricted use of specific sign language phenomena such as classifier signs. The distinguishing feature of the RWTH-PHOENIX-Weather corpus is that it uses real life data on a restricted domain and has a rather big size compared to other sign language corpora.

This paper is organized as follows: Section 2. discusses related work. The RWTH-PHOENIX-Weather corpus is described in Section 3.. Setups for sign language recognition and machine translation are defined in Section 4.. Preliminary recognition, tracking and translation results are presented in Section 5.. The paper is concluded in Section 6..

2. Related Work

The majority of freely available sign language video corpora are recorded under lab conditions and fullfill different purposes. "The American Sign Language Lexicon Video Dataset" (Athitsos et al., 2008), intended as a lexicon, was built with a vocabulary of 3000 signs but a total of only 3800 recorded glosses. The corpus provides gloss-level time alignment and front and side view video recordings. The AUSLAN project[1] collected 300 hours of Australian Sign Language in a similar approach. Further, there are linguistically motivated data sets with a more constrained vocabulary. The Corpus NGT (Crasborn and Zwitserlood, 2008) is a recording of 12 hours of signing in upperbody and front view and has 64 000 annotated glosses. Native signers varying in age and regional background sign in pair discussions or single sessions, using several variants of Dutch sign language without a specific domain. In a similar way, the sign language linguists at Boston University have published the freely available RWTH-BOSTON corpora (Dreuw et al., 2008) which comprise up to a total of 7 768 running glosses with a vocabulary size of 483.

[1]http://www.auslan.org.au

Figure 1: Example images and percentage of data performed by signer in RWTH-PHOENIX-Weather corpus. Top, left to right signers 1 to 4, bottom signers 5 to 7

(Braffort et al., 2010) presents corpora for isolated and continuous recognition in a front, side and closeup view. A motion tracker adds information to the manual annotations, and for a small part of the data non-manual gestures are provided.

The SIGNUM Database (von Agris et al., 2008), a data set aiming at the pattern recognition community, comprises 25 signers wearing dark clothes with long sleeves recorded in front view with a controlled background, signing a total of nearly 14 000 running glosses in German Sign Language (DGS) constrained to a vocabulary of 450 glosses. Based on an educational game for children, the CopyCat corpus (Zafrulla et al., 2010) comprises 59 different and a total of 420 ASL phrases based on a 19 word vocabulary signed by 5 deaf children.

3. Corpus / Database

Over a period of two years (2009 - 2010), the daily news broadcasts "Tagesschau" and "Heute-Journal" of the German public TV station "Phoenix" have been recorded. 190 weather forecasts of one and a half minutes have been annotated on the gloss level following the guidelines proposed by (Forster et al., 2010). Weather forecasting forms a rather compact domain in the sense that the vocabulary used is limited except for specific geographical references such as "Alps", "Berlin" or "Rhine".

Although the videos of the RWTH-PHOENIX-Weather corpus have not been recorded under lab conditions, the TV station tries to control the lighting conditions and the positioning of the signer in front of the camera. Additionally, signers wear dark clothes in front of a grey background. All videos have a resolution of 210 $\times$ 260 pixel and 25 frames per second (FPS). Figure 1 shows example images of all seven signers (six women and one man) present in the RWTH-PHOENIX-Weather corpus as well as the distribution of the signers in the overall corpus. One of the challenges of the RWTH-PHOENIX-Weather corpus is the signing speed, which in combination with the low temporal resolution of 25 FPS leads to motion blur effects. Furthermore, the signing itself has been performed by hearing interpreters under real-time constraints. This gives rise to two issues. First, the structure of the signing is closer to the grammatical structure of spoken German than in the case of other scenarios and second the signing features partly interrupted signs.

Figure 2: Pronunciation Variants of Sign GLATT (slippery)

The annotation has been performed using ELAN[2] and consists of

1. gloss sentences including sentence boundaries,
2. glosses including word boundaries,
3. pronunciation variants of glosses,
4. the utterances of the announcer in written German, annotated with the help of a speech recognition system,
5. an additional manual translation of the glosses into written German.

The annotation scheme used for the RWTH-PHOENIX-Weather corpus does not contain individual annotation tiers for each hand but assumes that the sign is performed with the right hand of the signer. Additional information about mouthings, content and interaction of the hands is provided in brackets following the glosses. Pronunciation variants have been annotated for Signers 01 and 03. A pronunciation has been considered distinct if the articulation differs strongly in the movement trajectory or the used hand-shape. Movement-epenthesis has been neglected for variant annotation. Figure 2 shows a pronunciation variant due to a different hand shape of the sign GLATT and Table 1 summarizes the major annotation conventions used in the RWTH-PHOENIX-Weather corpus.

The RWTH-PHOENIX-Weather corpus consists currently of 1 980 sentence in DGS, not counting annotations labeled as "<PAUSE>", and 22 822 running glosses. The overall vocabulary comprises 911 different signs. Table 2 summarizes the overall statistics of the RWTH-PHOENIX-Weather corpus.

To obtain the text spoken by the announcer, the open-source speech recognition system RASR (Rybach et al., 2009) was applied to the audio stream of the videos. The recognition output was then manually corrected by native German speakers to obtain the references used for the machine translation experiments. We found that this semi-automatic annotation scheme is fast and efficient. The corpus statistics for the translation corpus can be found in Table 2 in the right column. Note that the statistics differ

[2]http://www.lat-mpi.eu/tools/elan

Table 1: Annotation Scheme

Scheme	Example
gloss in capital letters	WIE-IMMER
finger spelling split by +	A+G+N+E+S
compound glosses split by +	V+LAND
numbers in written form	SIEBEN instead of 7
pointing gestures	IX
extended repetitions	SONNE++
pronunciation variants	TAG#1 TAG#2
classifier signs	cl-KOMMEN
lefthand only signs	lh-SONNE
signs negated by headshake	neg-WIND
signs negated by the alpha rule	negalp-MUESSEN
localization	loc-REGEN
additional mouthing	GLOSS-(mb:hill)
additional facial expression	GLOSS-(mk:strong)
additional localization	GLOSS-(loc:alps)
additional object of sign	GLOSS-(obj:cloud)

Table 2: Statistics of the RWTH-PHOENIX-Weather corpus for DGS and announcements in spoken German

	DGS	German
# signers	7	
# editions	190	
duration[h]	3.25	
# frames	293,077	
# sentences	1,980	2,640
# running glosses	21,822	32,858
vocabulary size	911	1,489
# singletons	537	525

somewhat from those presented for DGS. First of all, the glosses were annotated by a deaf expert, who also defined the segmentation of the gloss into sentences. The number of these segments does however not necessarily correspond to the number of spoken sentences. For the translation corpus, we therefore resegmented the glosses into sentences such that they correspond to the announcements in spoken German. Consequently, the number of segments differ.

In addition to the annotation of the signs in gloss notation and the announcements in written German, the center point of the hand palms and the nose tip have been annotated in a subset of 266 signs of the corpus. All seven signers are represented in the selected subset, which covers 39 691 frames. This subset can be used to evaluate hand and face tracking systems on real life data in the context of sign language and gesture recognition. The central image of Figure 3 shows an example of the tracking annotation. Furthermore, the Institute of Interactive and Itelligent Systems at the University of Innsbruck annotated 38 facial landmarks for all seven interpreters in a total of 369 images. These images cover the majority of variation in the facial expression and orientation of each signer and allow to train signer specific active appearance models. Examples of the facial annotation labeling are shown in Figure 3 on the left and right.

A first version including all data and a detailed description of the annotation schemes will be available on request from the Human Language and Pattern Recognition group at RWTH Aachen University starting in late June 2012. In the near future, we plan to release enhanced versions containing more data and additional annotations.

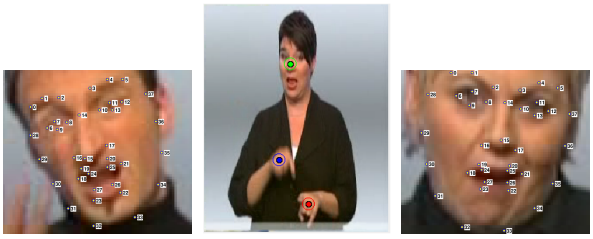

Figure 3: Visualization of Tracking (center) and Facial Annotations (left and right)

4. Evaluation Corpora

The RWTH-PHOENIX-Weather corpus allows for the creation and evaluation of automatic recognition systems for continuous and isolated sign language in a signer dependent, multi-signer and signer independent fashion as well as the creation and evaluation of automatic translations systems for the language pair DGS and German. Furthermore, the tracking and facial annotation allow for the evaluation of hand tracking and face tracking systems as well as face detection systems.

So far, a signer dependent subset for continuous sign language recognition, a setup for machine translation and a setup for the evaluation of hand and face tracking systems have been defined. A multi-signer subset for sign language recognition has been defined but is omitted here for brevity.

Table 3 shows the statistics for the signer-dependent setup. Signer 03 has been chosen in the signer dependent case because she covers more than 20% of all shows and is more consistent in the used signs than the other interpreters.

The signer-dependent setup allows to evaluate novel approaches to sign language recognition without facing the challenge of signing variations between different signers. Still, the signer-dependent setup forms a challenging task in itself with an out-of-vocabulary(OOV) rate of about 1.6% and a vocabulary of 266 signs. Out of these 266 signs, 90 appear only once in the training set. OOV signs cannot be recognized by an automatic sign language recognition system with closed vocabulary and singleton signs are only seen once during training making it difficult to train robust models. The signer-dependent task features intra-signer variation because the signers of the RWTH-PHOENIX-Weather corpus tend to mix different signing dialects. Furthermore, the signer-dependent task is a challenge with regard to computer vision techniques because of motion blur.

For the translation setup, the data was split into training, dev and test set as in the multi-signer recognition setup. However, as mentioned in the previous section, the segmentation of the glosses into sentences differs, since it is based on the segmentation of the spoken German sentences. Moreover, to simplify the complex annotation described in Table 1, we removed all additional annotations but included the information of mouthings and locations as additional

Table 3: Signer Dependent Subset for Signer 03

	Training	Test
# signers	1	1
# editions	41	31
duration [m]	31.09	4.5
# frames	46,638	6,751
# sentences	304	47
# running words	3,309	487
vocabulary size	266	-
# singeltons	90	-
OOV [%]	-	1.6

Table 4: Statistics of the RWTH-PHOENIX-Weather translation corpus

		Glosses	German
Train:	Sentences	2 162	
	Running Words	20 713	26 585
	Vocabulary	768	1 389
	Singletons/Voc	32.4%	36.4%
Dev:	Sentences	250	
	Running Words	2 573	3 293
	OOVs (running)	1.4%	1.9%
Test:	Sentences	2× 228	
	Running Words	2 163	2 980
	OOVs (running)	1.0%	1.5%

glosses (e.g., the gloss FLUSS-(loc:rhein) was transformed into FLUSS RHEIN). To further facilitate the statistical system to learn meaningful phrases from the scarce data, we introduced categories, mapping numbers, weekdays and months to special symbols. This enables the translation system to learn more general phrases. After the translation process, the categories are again replaced by their concrete values. Due to different preprocessing of the glosses described above, the vocabulary size on the gloss side differs from the statistics of the recognition corpus. For a summary of the corpus statistics of the translation setup, see Table 4. One issue of the RWTH-PHOENIX-Weather dataset is the rather loose translation by the sign language interpreters. Since the interpreters do not receive any transcripts in advance but have to interpret the announcements under real-time conditions, sometimes details are omitted in the sign language interpretation. This leads to an information mismatch in the corpus. To tackle this problem, we created additional translations of the gloss text into written German. When evaluating translations of the system, one can therefore take both the original announcements and the additional translation into account.

5. Preliminary Results

In this section, we present preliminary results using RASR, a statistical automatic speech recognition system, as well as video-based hand and head tracking system.

Table 5 shows tracking results obtained for the right (dominant) and left (non-dominant) hand as well as head tracking results obtained on the subset annotated for tracking evaluation. Hand tracking results have been achieved using a parameter free sliding window (50×70 pixel) approach in a dynamic programming framework (Dreuw et al., 2006). Head tracking results were obtained using a tracking by detection approach employing a generic project out active appearance model learnt using the facial annotations. The evalution has been carried out using the tracking error (TrackER) as the evaluation measure as defined by (Dreuw et al., 2006). Although the head tracking result is almost perfect, the hand tracking results suffer from motion blur, self-occlusion and crossing hands. TrackER measures the distance between the hypothesized object position and the groundtruth object position in pixel in the original resolution of the video. If the distance is larger then a previously defined threshold τ the current tracking hypothesis is counted as an error.

Table 5: Hand and Head Tracking Results

		TrackER			
Tracking	Model	τ=5	τ=10	τ=15	τ=20
Right Hand	no	82.9	51.1	25.1	11.6
Left Hand	no	88.2	63.4	41.7	28.9
Head	yes	6.44	0.69	0.27	0.15

Based on the tracking results for the dominant hand, cropped hand patch features have been generated and used for training and recognition for the signer-dependent setup. The RASR system achieves a Word Error Rate (Rybach et al., 2009) of 55.0 using a 3-gram language model underlining the challenge posed by the RWTH-PHOENIX-Weather corpus. The Word Error Rate measures the minimum number of insertion, deletion and substitution operations needed to transform a hypothesized string to the ground truth string.

For the translation of German Sign Language into spoken German, we used the RWTH in-house phrase-based statistical machine translation system. The glosses were preprocessed as described in Section 4., the German text was tokenized and lower cased. The word alignment was generated using GIZA++. [3] A detailed description of the methods applied for translation is given in (Stein et al., 2010). Note that here we use the same division of the data into train, dev and test as for multi-signer recognition setup so that future experiments can combine the sign language recognition and translation systems into a full pipeline. The translation results are presented in Table 6. BLEU (Papineni et al., 2002) is a metric based on the 4-gram precision and a brevity penalty, and thus higher values indicate a better translation quality. The translation edit rate (TER) (Snover et al., 2006) is an error metric counting the number of insertions, deletions, substitutions or shifts of whole phrases, and thus lower values indicate better translation results. The scores are calulated between the system output and one or several human reference translations.

In the first line, only the text spoken by the announcer was used as a reference. The score therefore measures how much the system translation resembles the announcements.

[3] http://www.hltpr.rwth-aachen.de/~och/software/GIZA++.html

Table 6: Translation Results DGS to German

Reference	BLEU	TER
Announcer	31.8	61.8
Multiple references	38.8	53.9

As mentioned before, there is some mismatch in information between the announcements and the signed utterances, as the sign language interpreter dropped some information or left out idioms specific to spoken German. In the second line, an additional translation which resembles the gloss text more closely was therefore added to the references. The improvements of 7 BLEU and 7.9 TER show that the additional references are helpful to alleviate the mismatch in the translations.

6. Conclusion

We introduced the RWTH-PHOENIX-Weather corpus, which is one of the largest freely available video based sign language corpora. The corpus is suitable for research in the area of sign language recognition and translation, and the additional annotations can be used in the areas of head and hand tracking as well as for the recognition of facial expressions and head orientations. Preliminary results have been presented as a baseline for further research. We hope that the corpus will be widely used by the scientific community in the area of sign language processing.

7. Acknowledgments

This work has been partly funded by the European Community's Seventh Framework Programme (FP7-ICT-2007-3. Cognitive Systems, Interaction, Robotics - STREP) under grant agreement n° 231424 - SignSpeak Project. Special thanks go to Philippe Dreuw, now with Bosch Research Hildesheim, and Daniel Stein, now with Fraunhofer IAIS Bonn, for their work and support.

8. References

V. Athitsos, C. Neidle, S. Sclaroff, J. Nash, A. Stefan, Q. Yuan, and A. Thangali. 2008. The american sign language lexicon video dataset. In *IEEE Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB)*, page 18, June.

A Braffort, L Bolot, E Chtelat-Pel, A Choisier, M Delorme, M Filhol, J Segouat, C Verrecchia, F Badin, and N Devos. 2010. Sign language corpora for analysis, processing and evaluation. In *LREC*, pages 453 – 456, Valetta, Malta, May.

O Crasborn and I Zwitserlood. 2008. The Corpus NGT: An Online Corpus for Professionals and Laymen. In Crasborn, Hanke, Efthimiou, Zwitserlood, and Thoutenhoofd, editors, *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages at LREC 2008*, pages 44–49, Paris. ELDA.

P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney. 2006. Tracking using dynamic programming for appearance-based sign language recognition. In *FG*, pages 293–298, Southampton, April.

P Dreuw, C Neidle, V Athitsos, S Sclaroff, and H Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *LREC*, Marrakech, Morocco, May.

J. Forster, D. Stein, E. Ormel, O. Crasborn, and H. Ney. 2010. Best practice for sign language data collections regarding the needs of data-driven recognition and translation. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies at LREC*, pages 92–97, Valletta, Malta, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July.

D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. 2009. The rwth aachen university open source speech recognition system. In *Interspeech*, pages 2111–2114, Brighton, UK, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA, August.

D. Stein, C. Schmidt, and H. Ney. 2010. Sign Language Machine Translation Overkill. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *IWSLT*, pages 337–344, Paris, France, December.

U. von Agris, M. Knorr, and K.-F. Kraiss. 2008. The significance of facial features for automatic sign language recognition. In *FG*, pages 1–6, Amsterdam, September.

Z. Zafrulla, H. Brashear, H. Hamilton, and T. Starner. 2010. A novel approach to american sign language (asl) phrase verification using reversed signing. In *CVPR workshops*, pages 48–55, San Francisco, CA, USA, August.