Best Practice for Sign Language Data Collections Regarding the Needs of Data-Driven Recognition and Translation

Jens Forster[1], Daniel Stein[1], Ellen Ormel[2], Onno Crasborn[2], and Hermann Ney[1]

[1]Human Language Technology and Pattern Recognition
RWTH Aachen University, Germany
{forster,stein,ney}@cs.rwth-aachen.de

[2]Department of Linguistics
Radboud University Nijmegen, Netherlands
{e.ormel,o.crasborn}@let.ru.nl

Abstract

We propose best practices for gloss annotation of sign languages taking into account the needs of data-driven approaches to recognition and translation of natural languages. Furthermore, we provide reference numbers for several technical aspects for the creation of new sign language data collections. Most available sign language data collections are of limited use to data-driven approaches, because they focus on rare sign language phenomena, or lack machine readable annotation schemes. Using a natural language processing point of view, we briefly discuss several sign language data collection, propose best practices for gloss annotation stemming from experience gained using two large scale sign language data collections, and derive reference numbers for several technical aspects from standard benchmark data collections for speech recognition and translation.

1. Introduction

Data-driven approaches to spoken language recognition and translation have seen great success over the last years. Common to all data-driven approaches is the need of large amounts of annotated data to learn reliable statistical models. In the case of natural languages, a data-driven system tries to learn individual statistical models for each phoneme respectively word requiring the system to see several utterances of a phoneme or word in the training data. Most sign language data collections have been created for linguistic research and as such tend to focus on rare phenomena. Both the focus on less frequent sign language phenomena and a low type-token ratio have so far limited the application of data-driven approaches to recognition and translation of sign languages.

Assuming the point of view of data-driven approaches, we briefly discuss the status of several sign language data collections in Section 2. and describe the needs of data-driven approaches to natural language processing in Section 3. Based on the status of the discussed sign language data collections, the needs of data-driven approaches, and experience gained in working with two large scale sign language data collections, we propose best practices for gloss annotation of sign language data collections. The practices proposed in Section 4. are designed for easy application to new and existing sign language data collection and allow for linguistic accurate annotation.

If a new sign language data collection is to be generated, the choice of the domain and some derived technical aspects like the targeted type-token-ratio and the vocabulary size are crucial variables that have a high impact on the performance of data-driven approaches. Based on existing data collection designed for speech recognition and translation, we provide reference numbers that can be used in the planning step for new data acquisition in Section 5.. The paper is concluded in Section 6.

2. Sign Language Data Collections

Although a full review of all available data collections is out of the scope of this work, almost all available data collections consist of annotated video material of various signers. The annotation has been typically conducted in glosses using specialized annotation tools such as ELAN[1], iLex (Hanke and Storz, 2008), or Signstream (Neidle et al., 2001). Gloss annotation assigns each sign the word from a spoken language that most appropriately describes the meaning of the sign. Besides the gloss annotation scheme, HamNoSys (Prillwitz et al., 1989) strives to describe signs on a phoneme-like level. All data collections discussed in this section have been annotated using glosses. Figure 1 shows example images taken from all data collections discussed in this work.

The RWTH-BOSTON (Dreuw et al., 2008) data collections are annotated subsets of data originally recorded at the Boston University. The annotations have been adjusted by RWTH Aachen University to fulfill the requirements of data-driven approaches. The data collections contain vocabulary sizes of up to 483 glosses, up to four different signers signing predefined sentences in front of a uniform background. Due to the small size of the data collections, and gray scale and color video recordings from lab environments, the RWTH-BOSTON data collections permit rapid development and testing of data-driven techniques for continuous sign language.

The ATIS (Bungeroth et al., 2008) data collection contains parallel annotation and videos for English, German, Irish sign language, German sign language, and South African sign language in the domain of the Air Travel Information System (ATIS). While the data collection can be used to build direct translation systems between different sign languages, the total size of only 600 parallel sentences is small in comparison to other sign language data collec-

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

tions. From a recognition point of view, the ATIS data collection contains challenging video recordings conditions including stark changes in illumination, cluttered office environments, and partial occlusions of the signer. In addition to the challenging recording conditions, the ATIS data collection shows for all included languages a singleton fraction of over 50%.

The European Cultural Heritage Online organization (ECHO)[2] published sign language data collections for Swedish sign language, British sign language, and sign language of the Netherlands (Crasborn et al., 2004). Although the data collection shows a high number of types, the chosen domain of fairy tales is challenging for data-driven approaches because of the intensive use of classifier signs.

Corpus NGT (Crasborn and Zwitserlood, 2008) is a large scale data collection for sign language of the Netherlands from several domains. Domains include fable stories, cartoon paraphrases, and discussions on sign language and Deaf issues. Especially the later two domains are interesting for data-driven approaches, because they allow for free discussions on topics with inherent limited vocabularies and hardly any classifier signs. Furthermore, sentence-aligned translations are currently created for the two discussion domains in the context of the EU funded SignSpeak project.

The SIGNUM data collection (von Agriss and Kraiss, 2008) has been specifically recorded for data-driven recognition of German sign language. The data collection contains over 700 predefined sentences signed by each of the 25 different native signers, and setups for signer dependent and signer independent recognition. The signers were asked to wear dark clothes and were recorded standing in front of a dark background.

Finally, the RWTH-PHOENIX data collection described by (Stein et al., 2010) contains German sign language for the domain weather forecast. The video material is recorded from broadcast news aired on the German television station Phoenix. Beside gloss annotation of the signs, translations into German are provided by a state-of-the-art speech recognition system for German. The chosen domain and employed annotation scheme are chosen with data-driven approaches in mind.

3. Needs of Statistical Recognition and Translation

Data-driven approaches to pattern recognition and model learning strive to learn a statistical model from the provided input data that best explains the input data in terms of a provided annotation. In the case of sign language recognition, the input data is a video stream showing a signing person with the annotation being the assigned gloss. For data driven translation, the input is a text in the source language e.g. glosses and the annotation is the corresponding text in the target language e.g. spoken language. Since data-driven approaches try to explain the input data in terms of statistical models, a system needs to collect several different examples of data labeled by the same annotation to incorporate the typical variance of the input data into the statistical model. Generally speaking, the more examples collected for a given annotation the better becomes the resulting statistical model.

In most cases the raw input data is difficult to explain by a statistical model due to high variance. Therefore, features are extracted from the raw data that allow for better discrimination between different annotations. The process of feature extraction strongly depends on the modality of the input data. While translation systems apply e.g. morphological parsing to the input data, vision-based recognition systems normalize the illumination, extract oriented gradients, and track the hands of the signer. Robust feature extraction in the presences of changing illumination, motion blur, scale changes, partial occlusion, and cluttered backgrounds is difficult to achieve using state-of-the-art computer vision techniques. Sign languages are especially prone to motion blur because of fast moving hands and abrupt motion changes. To ease the burden of feature extraction in video streams, we propose to limit the variability of the video streams by using standardized recording settings and high definition cameras capturing more than 30 frames per second.

Besides the statistical model explaining the input data, recognition and translation systems employ an additional knowledge source called the language model. The language model is learned from the annotations and assigns a probability to a sequence of annotations e.g. glosses based on the seen annotation sequences. Since the language model is learned from annotations, the language model depends on the domain of the annotations.

4. Best Practices for Gloss Annotation

Every variation in the annotation of a sign, though clearly identifiable by a human reader, will be treated as a new token by the computer. Minor concerns in the variation include spelling, capitalization, and linguistic comments within the annotations. While the first two minor issues can be enforced by the application of specific annotation parsers, linguistic comments contain additional information that cannot be extracted from a raw video stream. We propose to generally store all linguistic comments in a separate annotation or if you use ELAN a separate annotation tier.

4.1. Dialectic Signing Variants

A major issue in sign language annotation is the question of how to deal with dialectic signing variants. Dialectic signing variants of a word e.g. “MONDAY” are typically annotated by the same gloss in sign language data collections. However, dialectic variants of signs differ strongly in their appearance. If dialectic variants are annotated using the same gloss, a data-driven recognition system will learn a single model that tries to explain all dialectic variants of the sign in question. Ideally, each dialectic variant is represented by a distinct stochastic model that explains only this particular dialectic signing variant. To be able to train such a dialect specific model from data, the dialectic variants of a sign need to be consistently annotated by distinct glosses. Therefore, we propose to enumerate dialectic variants by applying the number as a postfix to the parent

[2] http://echo.mpiwg-berlin.mpg.de/home

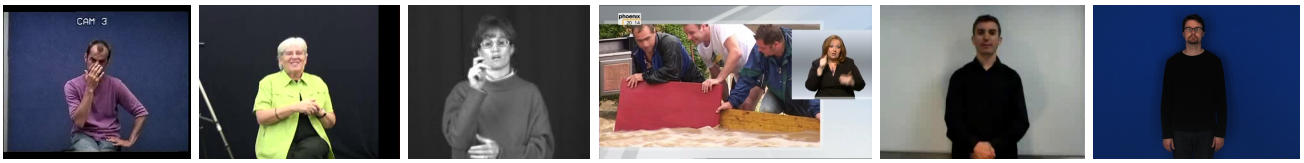

Figure 1: Example images from different sign language data collections (f.l.t.r.): ECHO, Corpus-NGT, RWTH-BOSTON, RWTH-PHOENIX, ATIS, and SIGNUM

gloss e.g. “MONDAY1”, “MONDAY2”, etc. This procedure has been applied in creating the RWTH-BOSTON data collections and is applied in the extension of the RWTH-PHOENIX and Corpus NGT data collections. In order to keep track of the numerous dialectic variants and to keep the annotation consistent, we propose to build a database containing video examples of dialectic variants of every gloss.

4.2. Homonyms and Synonyms

Related to the question of dialectic signing variants is the question of how to annotate homonyms and synonyms. Special to sign languages is the fact that there are true homonyms such as the sign for “DOCTOR” and “BATTERY” in sign language of the Netherlands and homonyms that share the same manual components but differ in mouthing. While true homonyms do not pose a problem to data-driven approaches as long as they there are consistently annotated by the same gloss and a list of true homonyms is provided to ease data-driven translation, the second class of homonyms, called *Umbrella-Glosses* in Corpus NGT, requires special care. An example of such an *Umbrella-Gloss* is “PROGRAMMA” which, depending on the mouthing, can mean rules or laws in sign language of the Netherlands. We propose to either split the annotation of the manual and non-manual parts of a sign into separate annotation files or tiers annotating e.g. “PROGRAMMA” for the manual part and “REGELS” for the non-manual part. As an alternative, we suggest to annotate an *Umbrella-Gloss* by its umbrella class followed by a delimiter and the actual realization of the umbrella. An example of the later approach is “PROGRAMMA:REGELS” and “PROGRAMMA:WETTEN” found in Corpus NGT. An advantage of the later approach is that a list of Umbrella-Glosses can be automatically generated from the annotation files.

In the case of synonyms, human annotators tend to use the meaning of a sign that is most appropriate in the context of the current sentence. By doing so, a synonym sign gets annotated by different glosses in one data collection effectively taking away observations from the model to be learned for the core meaning of this sign and biasing models for the synonym meanings. Consider for example the German signs for cathedral and carnival which are synonyms for Cologne and occur frequently in German broadcast news. We propose to use the glosses “CATHEDRAL” and “CARNIVAL” instead of Cologne and mark the intended meaning by an additional explicit postfix such as “-(syn:COLOGNE)”. Again we propose to generate a database containing video examples of synonyms.

4.3. Compound Glosses

Besides homonyms and synonyms, there exist several sequences of signs that need to be annotated by a single compound gloss to encompass its full meaning. An example is the gloss for gebarentaal (sign language in Dutch) that is composed of the sign “GEBAREN” followed by the sign for “TAAL” in sign language of the Netherlands. From a speech recognition point of view, the best procedure is to learn distinct models for “GEBAREN” and “TAAL” while from a translation point of view it is best to learn a model for “GEBARENTAAL”. To cope with this mismatch and facilitate accurate linguistic annotation, we propose to separate the glosses for “GEBAREN” and “TAAL” by a distinct delimiter such as ^ and to add the compound gloss “GEBARENTAAL” as additional information. This leads to a notation like “GEBAREN^TAAL:GEBARENTAAL” as it has been adopted for Corpus NGT. A similar notation will be used in the extended RWTH-PHOENIX data collection.

4.4. Finger Spelling

Finger spelling has an analog in word spelling for spoken languages. In the annotation of spoken languages spelled characters receive distinct annotations so that data-driven recognition and translation systems can learn distinct models for each spelled letter. For the sign language data collections discussed in Section 2., a sequence of finger spelled letters is often annotated as a single gloss such as “TREE” rendering it indistinguishable from a sign that does not employ finger spelling. Again, a data-driven system would try to learn a model for “TREE” although distinct models for each of the spelled letters would be more robust because the models are not only learned from the spelling of “TREE” but from all occurances of finger spelled letters. Therefore, we propose to either use distinct gloss annotations such as “T R E E” or to prefix finger spelled sequences by a special delimiter such as “#”. The later solution has the benefit that the amount of finger spelling in a data collection can be inferred automatically, it is less cumbersome to annotate, and existing annotations can be easily adapted to the proposed scheme.

4.5. Incorporation

Incorporation of signs is a common feature of sign languages. Typically two signs e.g. the sign for five and the sign for month are fused into a new sign featuring aspects of both parent signs. Since an incorporated sign is neither of the parent signs, a data-driven recognition system has to consider it a distinct class to be modelled from the given data. In order to distinguish the parent sign forms and incorporated sign form and to still keep information on the

parent signs, we propose to build the gloss annotation of the incorporated sign by connecting the glosses for the two parent signs by a hyphen e.g. "5-MONTH". This scheme has been employed in the RWTH-PHOENIX data collection and is successfully used in data-driven recognition.

4.6. Pointing and Referencing Signs

One of the strengths of visual languages is the possibility to refer to specific points in the signing space. Signers typically use pointing and referencing signs to convey temporal and causal concepts as well as relations between persons and objects. Except for self-referencing, the meaning of pointing and referencing signs is context dependent. The context of a referencing sign (e.g. the name of a person) can normally not be observed from the referencing sign itself. Since the context is known to the annotators, they typically use the context of a pointing or referencing sign to gloss such a sign. The information that a pointing or referencing sign has been used is lost. Therefore, it is difficult for a data-driven recognition or translation system to train robust models for pointing and referencing signs. Additionally, stochastic model used for the context of a pointing or referencing sign (e.g. the sign for "TREE") is biased by the visual content of the referencing sign. To limit the effects of annotating a pointing or referencing sign by its current context, we propose to include the context of a pointing or referencing sign as an additional information to the used gloss for pointing or referencing. As an example consider the notation adopted for Corpus NGT where a pointing/referencing sign is annotated by the gloss "IX" regardless of the intended context or e.g. consider the notation adopted for the RWTH-PHOENIX database where additionally to the gloss "IX" information on the spot in the signing space and the intended meaning is attached to the signing gloss as e.g. "-(loc:A,tree)".

4.7. Classifier Signs

A typical feature in signed languages are classifier signs capitalizing on the concept of free movement of the hands within the signing space. Classifier signs are non-lexicalized signs that show extreme variance in appearance and production. While it is already difficult for human experts to describe and annotate the exact meaning of a classifier sign, data-driven approaches are so far not able to cope with them. We propose to mark classifier signs by a special tag such as the @ sign or "<CLASSIFIER>" to be able to automatically extract all classifier signs from a data collection or to be able to create subsets of a data collection without classifier signs. Besides the information that a classifier sign has been used, it is desirable to add the perceived meaning of the classifier sign as additional information to the gloss marking. The proposed handling of classifier signs has been successfully used in our work with data from Corpus NGT.

4.8. Machine-Readability

Finally, all proposed practices for gloss annotation are useless to the natural language processing community if the annotation itself is not machine readable, consistent, accurate, and adequate. Machine readability is a prerequisite to automatic processing and parsing of large amounts of annotation data. This aspect includes the question of the used character encoding, preferably "UTF-8", and the choice of gloss delimiters. We propose to separate glosses by spaces and to avoid spaces within glosses and attached additional information. Further, we suggest to put additional information behind the relevant gloss annotation e.g. "GEBAREN^TAAL:GEBARENTAAL" and to use specific delimiters such as e.g. "^", "-", and ":" for different constructs as e.g. compound glosses or incorporation. In most annotation scenarios there will arise special cases requiring a special mark or prefix such as e.g. "@" or "#" to be applied to a gloss annotation. In such special cases, we propose to use unique marks not used in the remaining glossing scheme. The benefit of adhering to the proposed procedure is that the resulting annotation scheme is machine readable and can be automatically checked for consistency w.r.t. the chosen annotation scheme.

4.9. Adequacy of Annotation

Adequate annotation is crucial to data-driven systems because a data-driven system can only learn from data what can actually be seen in the data. For example, in most sign languages a negation of a sign is only conveyed by shaking the head parallel to performing the manual components of the sign. If a sign language recognition system is based on the manual components it will not be able to recognize the negation of a sign because the negation is only visible in the non-manual part. We suggest to split the annotation of manual and non-manual components such as eye gaze, shoulder movements, and facial expressions into distinct annotation files or tiers and to limit the annotation for each modality to what can actually be seen in the data for the modality in question at the given time. The proposed procedure eases the process of building specific statistical models for each modality and reduces errors in the systems. For data-driven translation, the parallel annotation of the glosses in another sign language or spoken language should be adequate in the sense that the glosses are translated as literally as possible without aiming for fluency in the target language. As an example a heavy nodding of the head accompanying the gloss "YES" we propose to translate by "yes, very much" rather than by "yes, I think this is a very good idea!".

5. New Data Collections

Independent of the chosen language, data collections of natural languages are hardly usable for data-driven approaches if the needs of data-driven approaches (cf. Section 3.) have not been taken into account when creating them. Using two small scale data collections for speech recognition and translation as references, we propose reference numbers for several technical aspects of sign language data collections.

Tables 1 and 2 show the statistics of small scale data collections used in speech recognition and translation. Although these data collections are by far bigger than anything we will see for several years to come in sign language data collections, they are among to the smallest data collections available for data-driven approaches to spoken language recognition and translation. The Verbmobil II corpus depicted in Table 1 contains spontaneous German

Table 1: Speech Recognition – Verbmobil II Corpus, German language, Domain of travel and booking

	Training	Evaluation
# sentences	36,015	1,081
# running words	701,000	14,000
vocab. size	10,157	–
audio data [h]	61.5	1.6

Table 2: Speech Translation – IWSLT 2005 Corpus, Chinese-English, Domain of travel and booking

	Training	Devel	Eval
# sentences	22,962	500	500
# running words Chinese	165,999	3,522	6,085
# running words English	218,829	62,517	54,22
vocab. size Chinese	8,786	948	1,328
vocab. size English	7,944	3,878	2,347

speech. The domain of the data collection is limited to travel and booking information, i.e. the data collection contains speech about how to get to Cologne by train but no information about sports. The IWSLT 2005 corpus shown in Table 2 is a bilingual translation data collection featuring parallel sentences in Chinese and English from the travel and booking domain. Both data collections have in common that they focus on the single domain of travel and booking. The focus on a single domain is preferable because current state-of-the-art speech recognition and translation systems use domain specific models. For new sign language data collections, we propose to consider domains in which classifier signs are less likely to occur.Although sign language recognition systems will overcome the problem of recognizing classifier signs over time, classifier signs will remain difficult to automatically recognize and translate over an extended period of time.

Besides the choice of the domain, the average type-token-ratio is a key technical aspect that should be considered when creating new sign language data collections. The Verbmobil II corpus shows an average type-token-ratio of 69.01, and the IWSLT 2005 corpus an average type-token-ratio 18.8 respectively 27.54 for Chinese respectively English. The high average type-token-ratio of the Verbmobil II corpus is special to this corpus and not normally found in data collections used in speech recognition. Although the higher the type-token-ratio the better for a data-driven system, a type-token-ratio of 69.01 will not be achievable for sign language data collections in a reasonable time frame. Other well-known standard data collections in speech recognition such as the Wall Street Journal data collections (Paul and Baker, 1992) typically have type-token-ratios between 15 and 40. Taking into account the needs of data-driven speech recognition and translation, one goal in the recording of new sign language data collections should be an average type-token ratio of about 20.

The average type-token-ratio as such is a misleading figure, because the average can be biased by a small number of very frequent tokens while the majority of tokens occurs only once or twice in a data collections. Therefore, the number of signs that occur only once in the data collection should be low. These singletons are in most cases named entities such as sign names or city names. In the Verbmobil II corpus and IWSLT 2005 data collections and several other benchmark databases for translation and speech recognition the percentage of singletons in the vocabulary is below 40%. This figure carries over to sign languages.

As already mentioned, the size of sign language data collections in terms of running signs or vocabulary size will not approach even the numbers given in Tables 1 or 2 over the next years. In order to keep the costs and time effort to create a sign language data collections that is also usable for data-driven approaches reasonable, we propose to aim for a vocabulary size that does not exceed $4,000$ glosses (i.e. half the vocabulary size of IWSLT 2005 Chinese). Taking into account a desired average type-token-ratio of about 20 the envisioned data collections contains at most $80,000$ running signs or 10% of Verbmobil II.

Data-driven translation systems typically exploit context information of words or complete phrases when translating a text from one language into another. The context used is typically limited to one sentence in order to limit computational cost. Therefore, data-driven translation translates one sentence of the source language e.g. sign language to an adequate sentence in the target language e.g. spoken language. This scheme requires bilingual sentence annotation as used in the IWSLT 2005 data collection. Unfortunately, the calculation of grammar inferred re-orderings of words is a computational expensive problem. Therefore, all used translation data collections limit the average sentence length to a range of 5 to 15 words in the source language. For a sign language data collection suitable for data-driven translation systems a similar bound should be used.

6. Conclusion

Most sign language data collection currently available for scientific research are of limited use to data-driven approaches to recognition and translation. We discussed the status of several sign language data collections available for scientific research from the point of view of data-driven speech recognition and translation. Based on the needs of data-driven approaches, we propose best practices for gloss annotation that ensure machine readable and adequate annotation of sign language while still allowing linguistically accurate annotation. Furthermore, we provide hard numbers for several technical aspects of data collections stemming from standard benchmark data collection of spoken languages. These hard numbers can act as references in the planning step for the creation of new sign language data collections.

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 .

8. References

J. Bungeroth, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way, and L. van Zijl. 2008. The ATIS Sign Language Corpus. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, 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.

O. Crasborn, E. van der Kooij, A. Nonhebel, and W. Emmerik. 2004. *ECHO Data Set for Sign Language of the Netherlands (NGT)*. Department of Linguistics, Radboud University Nijmegen.

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark Databases for Video-Based Automatic Sign Language Recognition. In *International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

T. Hanke and J. Storz. 2008. iLex — A database tool integrating sign language corpus linguistics and sign language lexicography. In *3rd Workshop on the Representation and Processing of Sign Languages at International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

C. Neidle, S. Sclaroff, and V. Athitsos. 2001. SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data. *Behavior Research Methods, Instruments, and Computers*, 3(33):311–320.

D. B. Paul and J. M. Baker. 1992. The Design of the Wall Street Journal-based CSR Corpus. In *DARPA SLS Workshop*, USA, February.

S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. 1989. *HamNoSys Version 2.0. Hamburg Notation System for Sign Languages. An introductory guide*, volume 5 of *International Studies on Sign Language and Communication of the Deaf*. Signum, Hamburg, Germany.

D. Stein, J. Forster, U. Zelle, P. Dreuw, and H. Ney. 2010. RWTH-Phoenix: Analysis of the German Sign Language Corpus. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Malta, May.

U. von Agriss and K.-F. Kraiss. 2008. Towards a Video Corpus for Signer-Independent Continuous Sign Language Recognition. In *8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, September.