Semi-automatic annotation of semantic relations in a Swiss German sign language lexicon

Ebling, Sarah ; Tissi, Katja ; Volk, Martin

Abstract: We propose an approach to semi-automatically obtaining semantic relations in Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS). We use a set of keywords including the gloss to represent each sign. We apply GermaNet, a lexicographic reference database for German annotated with semantic relations. The results show that approximately 60% of the semantic relations found for the German keywords associated with 9000 entries of a DSGS lexicon also apply for DSGS. We use the semantic relations to extract sub-types of the same type within the concept of double glossing (Konrad 2011). We were able to extract 53 sub-type pairs.

Originally published at:
Ebling, Sarah; Tissi, Katja; Volk, Martin (2012). Semi-automatic annotation of semantic relations in a Swiss German sign language lexicon. In: 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon, LREC 2012, Istanbul, 21 May 2012 - 27 May 2012, 31-36.

Semi-Automatic Annotation of Semantic Relations in a Swiss German Sign Language Lexicon

Sarah Ebling*, Katja Tissi, Martin Volk***

* Institute of Computational Linguistics, University of Zurich, Binzmühlestrasse 14, 8050 Zurich, Switzerland
{ebling,volk}@cl.uzh.ch
** University of Applied Sciences of Special Needs Education, Schaffhauserstrasse 239, 8050 Zurich, Switzerland
katja.tissi@hfh.ch

Abstract

We propose an approach to semi-automatically obtaining semantic relations in Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS). We use a set of keywords including the gloss to represent each sign. We apply GermaNet, a lexicographic reference database for German annotated with semantic relations. The results show that approximately 60% of the semantic relations found for the German keywords associated with 9000 entries of a DSGS lexicon also apply for DSGS. We use the semantic relations to extract sub-types of the same type within the concept of double glossing (Konrad 2011). We were able to extract 53 sub-type pairs.

Keywords: Sign language, Swiss German Sign Language, semantic relations, double glossing

1. Introduction

Sign language lexica have been annotated with various types of linguistic information in the past, including semantic relations. For example, Konrad (2011) reports work on the German Sign Language Corpus Project[1] in which the semantic relations synonymy and antonymy were annotated manually.[2] We present an approach to the semi-automatic annotation of semantic relations. We are compiling a German–Swiss German Sign Language corpus of train announcements as part of our efforts in building a machine translation system for this language pair. Departing from the semantic relations obtained, we also experiment with automatically obtaining sub-types that belong to the same type within the concept of double glossing (Konrad 2011, 145–155).

In Section 2 we briefly describe two common sign language notation systems, of which one (Section 2.1) is a coding system, and the other (Section 2.2) is a transcription system according to the typology of van der Hulst and Channon (2010). We then introduce Swiss German Sign Language and describe an existing lexicon for this language (Section 3). This is the lexicon that we will extend with the semantic relations extracted as a result of our approach. We describe our approach in Section 4. In Section 5 we introduce the typology of signs by Johnston and Schembri (1999), which is the basis of the concept of double glossing. Double glossing is described in Section 6, where we also present our approach to automatically extracting sub-types of the same type.

2. Sign Language Notation Systems

2.1. Gloss Notation

Sign language glosses are semantic representations of signs that usually take the default form of the corresponding spoken language word.[3] For example, in Swiss German Sign Language (cf. Section 3), the gloss GESCHWISTER, a German word, is used to represent the sign for 'siblings'. Glosses can also consist of multiple words, e.g., SICH-SETZEN ('to have a seat').[4]

Glosses allow for alphabetic sorting in a lexicon. However, from a conceptual point of view it is problematic to express the vocabulary of one language (i.e., a sign language) by means of another (i.e., a spoken language). A further problem with glosses is that they are not standardized; the same sign may be denoted with multiple glosses. Moreover, glosses typically convey only limited facial expression and information about body movement. This means that they cannot, e.g., differentiate between different movement paths of the hands through which a signer associates different objects with individual locations in the signing space (Huenerfauth, 2006).

These shortcomings imply that glosses are merely sufficient to refer to entries in a lexicon. For all other purposes, e.g., for investigating the sublexical components of a sign, a more powerful notation system is needed. The Hamburg Notation System for Sign Languages (HamNoSys) (Prillwitz et al., 1989) has been developed for this.

2.2. HamNoSys

HamNoSys consists of approximately 200 symbols. It takes explicit account of sublexical components: each of the components handform, hand position (with extended finger direction and palm orientation as subcomponents), location, and movement is transcribed. Figure 1 shows the example of the sign NATION, VOLK ('nation', 'people') in Swiss German Sign Language that contains one instance of each component.[5]

[1] `http://www.sign-lang.uni-hamburg.de/dgs-korpus/`

[2] This was done based on relations between the underlying images of signs, i.e., based on iconicity.

[3] By 'spoken language' we mean a language that is not a sign language.

[4] We follow the convention of writing glosses in all caps.

[5] A sign generally consists of at most two syllables, with the maximum syllable represented as Hold–Movement–Hold, as in the sign BASEL ('Bâle') in Swiss German Sign Language.

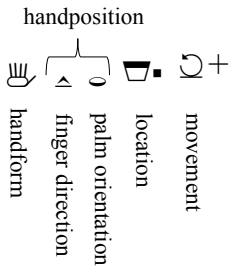


Figure 1: HamNoSys sublexical components for the sign NATION, VOLK (‘nation’, ‘people’) in Swiss German Sign Language

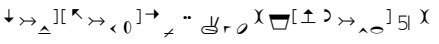


Figure 2: HamNoSys transcription of GEBÄRDENSPRACHKURS (‘sign language course’) in Swiss German Sign Language

HamNoSys is machine-readable: it offers an XML representation, the Signing Gesture Markup Language (SiGML) (Elliott et al., 2000), which can be used to drive an avatar.
The most significant disadvantage of HamNoSys is that its inventory does not include symbols to encode non-manual features. A set of two-letter tags was introduced in the ViSiCAST project to capture aspects like shoulder, body, and head movement as well as eye gaze and facial expression (Hanke, 2001). However, these tags are not officially part of the HamNoSys symbol set. Moreover, HamNoSys transcriptions are rather complex, as can be seen from Example 3, the transcription of the sign GEBÄRDENSPRACHKURS (‘sign language course’) in Swiss German Sign Language. Nevertheless, HamNoSys is considered by many to be the state-of-the-art sign language notation system. It is also used in a lexicon of Swiss German Sign Language. In the following section, we describe Swiss German Sign Language in more detail and introduce the lexicon.

3. Swiss German Sign Language

Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS) is the sign language of the German-speaking area in Switzerland. It has approximately 6000 users (Lewis, 2009) distributed across five different dialects. We focus on the Zurich dialect. The mouthings used in DSGS are derived from Standard German rather than from one of the Swiss German dialects. In 1996, work started on a DSGS lexicon (Boyes Braem, 2001). Currently the lexicon contains about 9000 signs. Each sign is represented with a German gloss as well as with a set of German keywords. Keywords are included since glosses often reflect only one meaning of the sign or do not sufficiently distinguish the sign’s meaning from the meanings of other signs. For example, there are two glosses GLAUBEN: one () carries the keywords *glauben*, *annehmen* (‘believe’, ‘assume’), the other () the keywords *glauben*, *Glaube*, *gläubig*, *religiös* (‘believe’, ‘faith’, ‘religious’).
In the lexicon, a video clip is available for the citation form of the sign. The possible modifications (e.g., pluralisation, verbal aspect) as well as additional linguistic information are given in textual form. Efforts are made to transcribe all of the signs in HamNoSys.
Our aim is to enrich the DSGS lexicon with information on semantic relations. We identify semantic relations in a two-step process: first, a set of pairs of semantically related entries is generated automatically; second, the set is filtered through manual screening by a native signer. In what follows, we describe the first step of this process in more detail.

4. Identifying Semantic Relations

4.1. Step 1: Automatic Identification

Relation	Word Class			Reverse Rel.	Type
	N	A	V		
hyperonymy	✓	✓	✓	hyponymy	conceptual
hyponymy	✓	✓	✓	hyperonymy	conceptual
meronymy	✓	✗	✗	holonymy	conceptual
holonymy	✓	✗	✗	meronymy	conceptual
entailment	✗	✗	✓		conceptual
causation	✗	✗	✓		conceptual
association	✓	✓	✓		conceptual
synonymy	✓	✓	✓	synonymy	lexical
antonymy	✓	✓	✓	antonymy	lexical
pertonymy	✓	✓	✓		lexical
participle	✗	✓	✗		lexical

Table 1: Sub-types of semantic relations in GermaNet

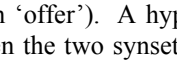
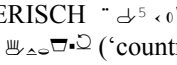
We use the sign language glosses and keywords available for each sign in the DSGS lexicon (cf. Section 3) as indicators of the underlying semantic concepts. We search for these words in GermaNet. GermaNet is a lexicographic reference database for German word senses containing annotations of semantic relations (Hamp and Feldweg, 1997). It is based on the Princeton WordNet for English (Miller, 1993).[6] For the semantic relations in GermaNet to be fully applicable to the entries in the DSGS lexicon, the semantic networks of the two languages DSGS and German would have to be exactly the same. This is not the case; yet we expect the networks to be similar enough for us to use the assumption of concept equality as a heuristic. We believe it is promising to further investigate the applicability of spoken language concepts to sign languages.
The basic units in GermaNet are semantic concepts, which are called *synsets*. Each synset combines a set of lexical units by which the corresponding concept is denoted. Lexical units may take multiple orthographic forms, e.g., *fantastisch* (‘fantastic’) may appear as *fantastisch* or *phantastisch* (new and old spelling, respectively). Lemmas can belong to one of the three word classes noun, verb, and adjective.

(1) `<synset id="s2376" category="adj">`

[6] Wordnets also exist for other languages, including Dutch, Italian, Spanish, German, and French (Vossen, 2004).

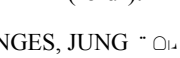
```
<lexUnit id="l3637" sense="1" source="core"
namedEntity="no" artificial="no"
styleMarking="no">
<orthForm>erforscht</orthForm>
</lexUnit>
<lexUnit id="l3638" sense="1" source="core"
namedEntity="no" artificial="no"
styleMarking="no">
<orthForm>erkundet</orthForm>
</lexUnit>
</synset>
```

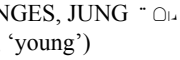
There are two main types of semantic relations in GermaNet: *conceptual relations* hold between synsets, and *lexical relations* hold between lexical units. Table 1 lists the sub-types of semantic relations available in GermaNet: for each relation, the word class(es) with which it occurs (N–noun, A–adjective, V–verb), the reverse relation (where available), and its type (conceptual or lexical) are specified. Example 1 shows the synset `s2376` containing the two lexical units `l3637` and `l3638` that are each represented by a single orthographic form: *erforscht* ('investigated'), and *erkundet* ('explored'). Synonym relations exist between the lexical units.

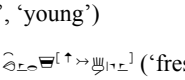
We use version 6.0 of GermaNet, which contains 93,407 lexical units distributed across 69,594 synsets. 81,852 conceptual relations hold between the synsets, and 3562 lexical relations exist between the lexical units. For each entry pair in the DSGS lexicon we check whether the glosses themselves or one of the keywords on each side are present in the GermaNet database. If this is the case, we extract the set of relations that exist between the lexical units under consideration or between the synsets to which the lexical units belong. For example, a synonym relation exists between the two entries ANGEBOT  and ANTRAG (both 'offer'). A hyponym/hyperonym relation exists between the two synsets to which the entries ALGERIEN, ALGERISCH  ('Algeria, Algerian') and LAND ('country') belong.

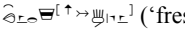
Sub-type	Absolute count	Percentage
Hyponym/hyperonym	5435	77.87
Synonym	817	11.70
Meronym	279	4.00
Antonym	277	3.97
Pertainym	59	0.85
Participle	40	0.57
Related-to	39	0.56
Causation	34	0.48

Table 2: Relation sub-types extracted along with their frequencies

We identified 6980 relations. The distribution according to relation sub-types is as shown in Table 2. Below is a sample output of our approach (Examples 2 to 7): six antonyms of the sign ALT  ('old').

(2) JUNGTIER, JUNGES, JUNG  ('young animal', 'young')

(3) FRISCH, NEU  ('fresh', 'new')

(4) JUNG, JUGENDLICHER, JUGEND  ('young', 'youth')

(5) KLEIN, JUNG, KLEINES, JUNGES ('small', 'young')

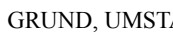
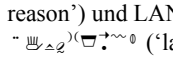
(6) NEU, BRANDNEU ('new', 'brand new')

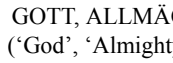
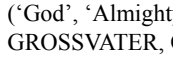
(7) NEU ('new')

4.2. Step 2: Manual Screening

Step 2 of our process consists of manually filtering the semantic relations that were retrieved automatically during Step 1. This task was carried out by a native signer who is also a member of our project. We presented her with 500 randomly selected statements of the kind displayed in Examples 8 and 9 and asked her to rate them with True or False. She rated 302 out of 500 statements with True (60.4%) and 198 statements with False (39.6%).

(8) LAND  ('country') is hypernym of ALGERIEN, ALGERISCH  ('Algeria, Algerian') ✓

(9) GRUND, UMSTAND, MOTIV  ('cause, reason') und LANDSCHAFT, UMGEBUNG, GEGEND  ('landscape, neighbourhood') have the same meaning ✗

(10) GOTT, ALLMÄCHTIGER, HERR, VATER  ('God', 'Almighty', 'Lord', 'Holy' 'Father') is hyponym of GROSSVATER, OPA, GROSSPAPA  ('grandfather', 'granddad').

The statements rated as False are relations that do not apply for DSGS. There are two possible reasons for this:

1. Our system found the relation at hand based on a sense of a German keyword that was incorrect in the given context. Hence, the relation is also not valid for German. For example, the (false) statement shown in Example 10 is due to an ambiguity of the word *Vater*, which can mean both 'Holy Father' as well as 'father'. In this case, the former ('Holy Father') is the intended meaning, whereas the hyponym GROSSVATER, OPA, GROSSPAPA ('grandfather', 'granddad') proposed by our system is based on the latter meaning ('father'), which in this context is incorrect.

2. The relation is valid for German but not for DSGS. Hence, it accounts for a difference in the semantic concepts of German and DSGS. For example, in German, *Künstler, Künstlerin, Kunstschaffender, Kunstschaffende* ('artist') is a hypernym of *Musiker, Musikerin, Musikant, Musikantin* ('musician'). In DSGS, however, KÜNSTLER, KÜNSTLERIN, KUNSTSCHAFFENDER, KUNSTSCHAFFENDE is restricted to the meaning of a visual artist, i.e., a painter. Hence, the relation does not hold for DSGS. Similarly, TRAINING, TRAINIEREN ('practice') is confined to the domain of physical activity in DSGS, whereas it may involve any sort of training in German. Hence, *Training, trainieren* is a valid hyponym of *Lehren, unterrichten, schulen, belehren, erklären* ('instruct', 'teach') in German but not in DSGS. As a third example, *Haushalt,*

haushalten (‘household’) is a hypernym of *Buchhaltung, buchhalten* (‘book-keeping’) in German but not in DSGS, where BUCHHALTUNG, BUCHHALTEN cannot have a financial aspect to its meaning (as in ‘financial household’).

We will incorporate the semantic relations identified as valid for DSGS into the DSGS lexicon (cf. Section 3).

Semantic relations between signs can be used to obtain sub-types that belong to the same type within the concept of *double glossing* (Konrad, 2011, 145–155).[7] Double glossing builds on the typology of signs by Johnston and Schembri (1999). In what follows, we introduce this typology. In Section 6.1 we then explain the concept of double glossing. In Section 6.2 we present our approach to automatically generating sub-types of the same type.

5. Typology of Signs

Johnston and Schembri (1999) assume two main types of signs in sign languages: productive signs and conventional signs.[8] For productive signs, the sum of meanings of the sublexical components (cf. Section 2.2) yields the overall meaning of the sign. Productive signs are iconic: their form can be traced back to an underlying image. Moreover, they are context-dependent. They are derived spontaneously, which is why they are also referred to as “on-the-spot” signs. It is for this reason also that productive signs do not have a stable citation form; hence, they do not appear in the lexicon.

Conventional signs are idiomatic in that their overall meaning is not composed merely of the meanings of the sublexical components. They are “off-the-shelf” signs, which makes them similar to spoken-language words. Most conventional signs are originally iconic, yet have developed into self-contained form-meaning units that can be used without drawing on iconic value. Nevertheless, the iconic value of many conventional signs can be reactivated by modifying the sign, e.g., by pluralising it. This process is called delexicalisation, or reiconisation. By contrast, the process during which productive signs are turned into conventional signs is referred to as lexicalisation.

Konrad (2011) extends the scheme of Johnston and Schembri (1999) by further differentiating conventional signs according to two usages: a conventional usage, and a productive usage. The usages are characterized by the way in which signs are combined with mouthings: a conventional usage of a conventional sign implies a habitual combination of sign and mouthing, while a productive usage is constituted by an unusual or occasional sign-mouthing combination.

Figure 3 summarizes the typology of signs proposed by Johnston and Schembri (1999) and extended by Konrad (2011). It is important to note that the distinction between productive and conventional signs is one of degree rather than kind. The concept of double glossing applies to conventional signs only. This is the part enclosed in the dotted frame in Figure 3.

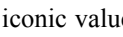
6. Double Glossing

6.1. Concept Overview

The concept of double glossing includes a two-level hierarchy: the upper level consists of (lexical) types, the lower level consists of sub-types that are identical in form, underlying image, and image producing technique.[9] Identical images and image producing techniques imply that signs have the same iconic value. Hence, the governing principle behind double glossing is iconicity. As an example, consider the type FLACH (‘flat’), which unites, among others, the sub-types BASIS (‘base’), BODEN (‘ground’), FELD (‘field’), and TISCH (‘table’) (Konrad, 2011). The common underlying image here is that of a flat surface. Other examples of type names are VIERECK (‘square’), BEREICH (‘area’), BEHÄLTER (‘container’), RUND (‘round shape’), and MATERIAL (‘material’) (Konrad, 2011).

6.2. Identifying Sub-Types from Semantic Relations

Departing from the semantic relations obtained from our two-step process (cf. Section 4), our goal was to automatically extract sub-types of the same type within the concept of double glossing. Hence, we were concerned with semantically related sub-types. This is the area outlined in grey in Figure 4. Outlined in black is the area in which double glossing as a whole takes place.

We looked at pairs of hyponyms of the same hypernym that are form-equivalent, i.e., have the same HamNoSys notation. For example, BACH (‘brook’) and FLUSS (‘river’) have the same form () and the same GermaNet hypernym, *Gewässer* (‘stretch of water’). The heuristic here is that two form-equivalent signs that are semantically related have the same underlying image (in this case, a tracing movement sketching the flow of a river or a brook, cf. Figure 5) and the same image producing technique. Note that this is not always true. Konrad (2011, 175) names the example of two different signs KRIPPE that are equivalent in their underlying image but not equivalent in form and image producing technique.

We identified 53 sub-type pairs of this kind. Note that it would be conceptually wrong to name the corresponding types after the mutual hypernyms, e.g., to choose GEWÄSSER as the name of the type comprising the sub-types BACH and FLUSS; this is because not all sub-types of a type are necessarily semantically related.

7. Conclusion

In this paper we showed that using a spoken-language resource to obtain information about a sign language leads to 60% correct semantic relations. We used GermaNet to automatically obtain candidates for semantic relations in Swiss

[7] Efforts are underway to include double glossing in the DSGS lexicon. Double glossing is implemented in iLex (Hanke and Storz, 2008).

[8] A third type exists that subsumes various smaller categories, e.g., finger alphabet, initialised signs, index signs, number signs, gestures, etc.

[9] The terms *type* and *sub-type* are from König et al. (2010). Note that they are different from the terms of the same name introduced in the context of semantic relations in Section 4. Image producing techniques are described in Langer (2005).

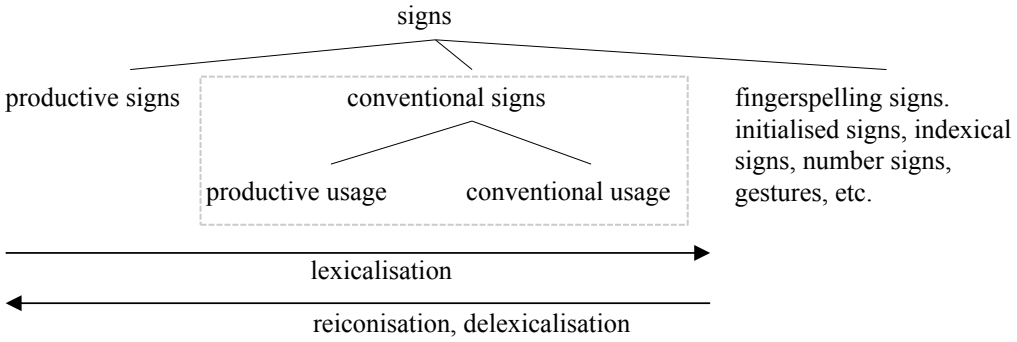


Figure 3: Typology of signs based on Johnston and Schembri (1999) and Konrad (2011)

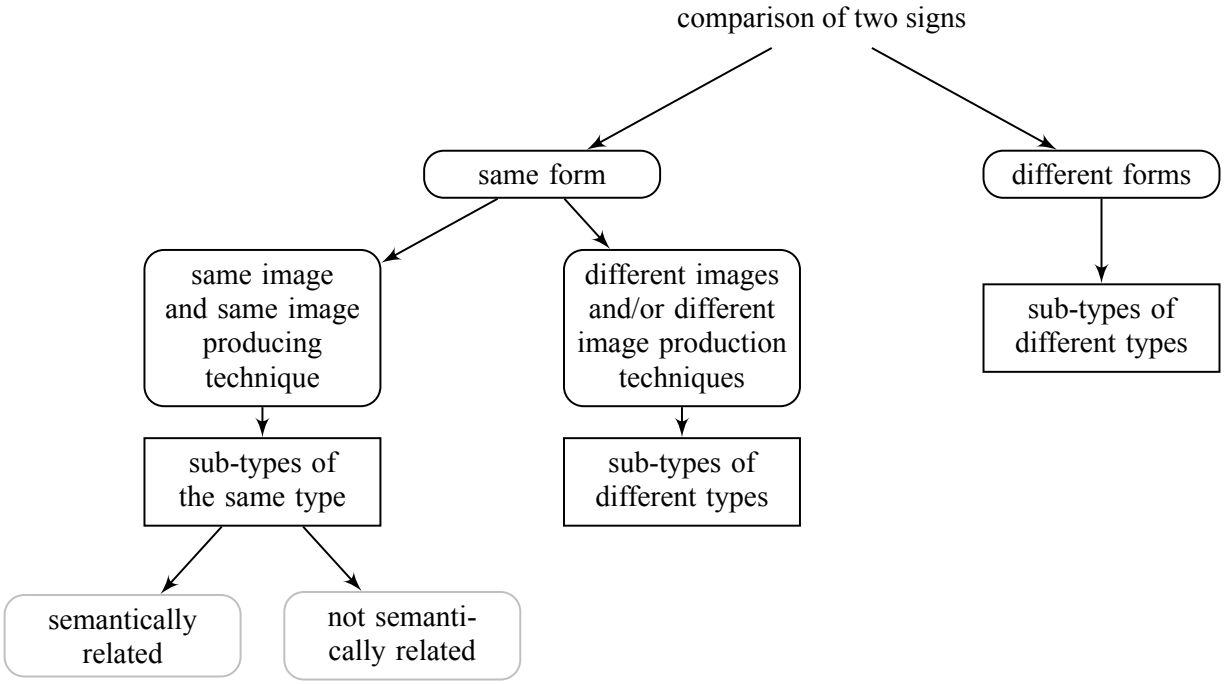


Figure 4: Double glossing: Identifying sub-types of the same type (based on Konrad (2011))

Figure 5: Form of the signs BACH ('brook'), FLUSS ('river'), and WEG ('path') (source: DSGS lexicon)

German Sign Language. We also looked at semantically related sub-types within the concept of double glossing. We extracted 53 sub-type pairs. Our approach contributes to a comparison of semantic and iconic networks: it yields further insight into the question raised by Konrad (2011, 238), "ob und inwieweit ikonische und semantische Netzwerke zur Deckung gebracht werden können" ('whether and, if so, how iconic and semantic networks can be brought to overlap').

In the future, we would like to look into ways of operationalizing the criteria of 'identical image' and 'identical image producing technique' so as to be able to extract more subtypes automatically. Given the high cost of double glossing (Konrad, 2011, 151) it seems reasonable to automate as much of this task as possible.

We will also investigate the possibility of extracting additional semantic relations for German from Swiss German Sign Language form equivalences. In doing so, we will pursue the opposite direction to that commonly investigated, i.e., we will attempt to arrive at additional knowledge of

spoken languages using information from sign languages.

8. Acknowledgements

We thank Penny Boyes Braem for providing her DSGS lexicon as well as giving advice at various stages of this work, and Reiner Konrad for his helpful comments.

9. References

Penny Boyes Braem. 2001. A multimedia bilingual database for the lexicon of Swiss German Sign Language. *Sign Language & Linguistics*, 4:133–143.

R. Elliott, J. R. W. Glauert, J. R. Kennaway, and I. Marshall. 2000. The development of language processing support for the ViSiCAST project. In *Proceedings of the fourth international ACM conference on Assistive technologies*, Assets '00, pages 101–108, Arlington, Virginia, United States. ACM.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid.

Thomas Hanke and Jakob Storz. 2008. iLex – A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *LREC 2008 Workshop Proceedings*, pages 64–67.

Thomas Hanke. 2001. ViSiCAST Deliverable D5-1: Interface Definitions. Technical report, ViSiCAST project.

Matt Huenerfauth. 2006. *Generating American Sign Language Classifier Predicates for English-to-ASL Machine Translation*. Ph.D. thesis, University of Pennsylvania.

Trevor Johnston and Adam Schembri. 1999. On Defining Lexeme in a Signed Language. *Sign Language & Linguistics*, 2:115–185.

Susanne König, Reiner Konrad, Gabriele Langer, and Rie Nishio. How much top-down and bottom-up do we need to build a lemmatized corpus? Poster presented at the *Theoretical Issues in Sign Language Research Conference (TISLR 10)*, Sept 30–Oct 2, 2010, Purdue University, Indiana, USA. Online: `http://www.purdue.edu/tislr10/pdfs/Konig%20Konrad%20Langer%20Nishio.pdf`.

Reiner Konrad. 2011. *Die lexikalische Struktur der Deutschen Gebärdensprache im Spiegel empirischer Fachgebärdenlexikographie. Zur Integration der Ikonizität in ein korpusbasiertes Lexikonmodell*. Gunter Narr Verlag.

Gabriele Langer. 2005. Bilderzeugungstechniken in der Deutschen Gebärdensprache. *Das Zeichen*, 70:254–270.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Tex., 16th edition. Online version: `http://www.ethnologue.com/`.

George A. Miller. 1993. WordNet: a Lexical Database for English. In *Human Language Technology, Proceedings of a workshop held at Plainsboro, New Jersey*, pages 409–409, Plainsboro, New Jersey.

Siegmund Prillwitz, Regina Leven, Heiko Zienert, Thomas Hanke, and Jan Henning. 1989. *HamNoSys: Version 2.0: Hamburg Notation System for Sign Languages. An introductory guide.* Signum, Hamburg.

Harry van der Hulst and Rachel Channon. 2010. Notation systems. In Diane Brentari, editor, *Sign languages*, Cambridge language surveys, pages 151–173. Cambridge University Press, Cambridge.

Piek Vossen. 2004. EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17:161–173.