Glossing a multi-purpose sign language corpus

Ellen Ormel[1], Onno Crasborn[1], Els van der Kooij[1], Lianne van Dijken[1], Yassine Ellen Nauta[1], Jens Forster[2] & Daniel Stein[2]

[1] Centre for Language Studies, Radboud University Nijmegen, PO box 9103, NL-6500 HD Nijmegen, The Netherlands

[2] Human Language Techinology and Pattern Recognition, RWTH Aachen University, Germany

E-mail: e.ormel@let.ru.nl, o.crasborn@let.ru.nl, e.van.der.kooij@let.ru.nl, l.vandijken@let.ru.nl, e.nauta@let.ru.nl, forster@i6.informatik.rwth-aachen.de, stein@i6.informatik.rwth-aachen.de

Abstract

This paper describes the strategies that have been developed for creating consistent gloss annotations in the latest update to the Corpus NGT. Although the project aims to embrace the plea for ID-glosses in Johnston (2008), there is no reference lexicon that could be used in the creation of the annotations. An idiosyncratic strategy was developed that involved the creation of a temporary 'glossing lexicon', which includes conventions for distinguishing regional and other variants, true and apparent homonymy, and other difficulties that are specifically related to the glossing of two-handed simultaneous constructions on different tiers.

1. Introduction

Over the past years, various initiatives in the area of signed language annotation have been undertaken, but in the area of sign language glossing, no clear standards have been developed (Schembri & Crasborn, this volume). To some extent, researchers lean towards the general principles of the Leipzig Glossing Rules[1], but these do not specifically mention sign language data and the concomitant challenges. An important contribution to the discussion has been Johnston's (2008) emphasis on the use on 'ID-glosses': identical forms should be consistently glossed, and variant forms should receive distinctive glosses.

Work on corpus construction, including the creation of annotations, has recently been increasing and is currently carried out for different sign languages other than Sign Language of the Netherlands (NGT), for example, for Auslan (e.g., Johnston, 2008; Johnston, 2009; Johnston, Vermeerbergen, Schembri, & Leeson, 2007), British Sign Language (BSL, e.g., Schembri, 2008,), and German Sign Language (DGS, e.g. Hanke, 2002; Hanke, Konrad, & Schwarz, 2001).

Machine processing of signed languages has become an active research field as well, testified by the LREC workshop series. In order to facilitate machine processing of sign language corpora, several points need to be considered. In the present paper, we describe some of the adaptations in the Corpus NGT in order to facilitate machine processing.

A specific problem in the creation of the Corpus NGT was the absence of a lexicon with unique lemmata and variants that could be referred to. The dictionaries that have been published in the Netherlands are fragmented in focussing either on basic lexicon or on specific topics. In the last ten years, dictionary products have explicitly excluded variation with the aim of promoting standardisation of the lexicon (Schermer, 2003; Crasborn & Bloem, 2009; Crasborn & de Wit, 2005).

This paper will discuss the process of finding gloss conventions for the Corpus NGT that on the one hand function like ID-glosses, and on the other hand can be consistently created in the absence of a reference lexicon.

2. First release of the glossing conventions of the Corpus NGT

2.1 The Corpus NGT

The first release of the Corpus NGT in 2008 was created in a two-year project funded by the Netherlands Organisation for Scientific Research (NWO, grant no. 380-70-008), aimed at collecting a set of data from deaf signers using NGT (Crasborn, Zwitserlood & Ros, 2008, Crasborn & Zwitserlood, 2008). It has been completed in 2008, with the publication of the corpus on Internet.[2] The data consist of recordings with multiple synchronised video cameras, accompanied by gloss and translation annotations. All data are freely accessible to researchers and the general public. In each corpus video, a maximum of two subjects participated (S1 and S2). The left hand is glossed separately from the right hand.

All annotations were created in the ELAN software[3] (see also Crasborn & Sloetjes 2008, 2010). This annotation tool allows multiple annotation layers ('tiers') to be time-aligned with several video files (Figure 1).

Every annotated file contains the following tiers: *GlosL S1, GlosR S1, GlosL S2, and GlosR S2*. These four tiers contain the glosses for the activities of the left hand (GlosL) and the right hand (GlosR) respectively, of the signer to the left (S1) and the signer to the right (S2). In signs made with two hands, the hands do not always move precisely simultaneously (Figure 2). Often, one hand stays in the final position of the sign, while the other hand starts articulating the next sign. Or one hand starts slightly earlier than the other hand. For each hand, the precise

[1] http://www.eva.mpg.de/lingua/resources/glossing-rules.php.

[2] http://www.ru.nl/corpusngtuk

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

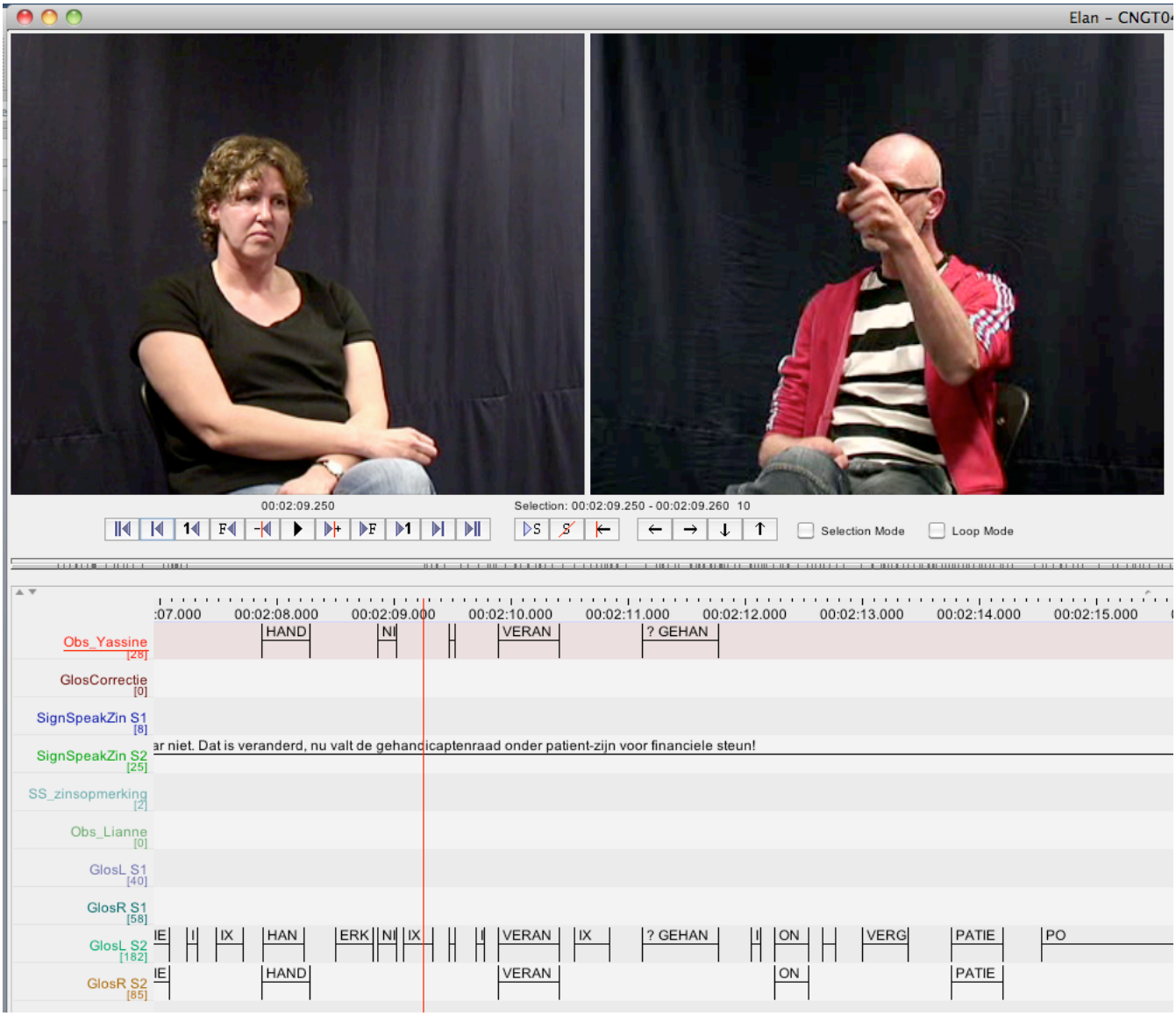

Figure 1. Multi-tiered annotation of multiple video files in ELAN

duration of the presence of a sign is shown in the gloss on the GlosL- or GlosR-tier.

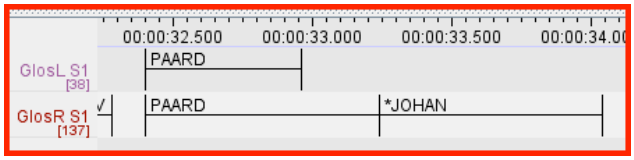

Figure 2. Time alignment of glosses per hand

2.2 Initial glossing conventions

The glosses in the annotation files in the *Corpus NGT* are intended to indicate the exact start and end time of the signs, as well as to refer to a lexicon. Thus, the glosses in Dutch are not actual translations; in the ideal case they are pointers to lemmas in a lexicon. Because of the fact that there is no common orthography for sign language or a practical, commonly used phonetic notation system (Miller, 2001), Dutch words have been used as a reference, rather than first glossing in the language itself and subsequently translating the glosses to English or Dutch for accessibility, as is more commonly done for spoken languages. The Dutch glosses that are used approach (one of) the meaning(s) of the signs; however, the real meanings of the sign forms are described in the lexicon, not by the gloss. Exceptions to this rule are non-lexicalized forms that, in the gloss, are preceded by a @-character (see under #4 below).

Although it was our intention to use glosses referring to a lexicon, it was impossible to always consult the lexicons

of the Dutch Sign Centre (NGc) on DVD or on the internet, given the way the glosses were established and for reasons of efficiency. Because of this, the glosses in the first release contained many inconsistencies that users of the annotations need to be aware of. It is expected that many files contain a number of inconsistencies as well as interpretation differences and mistakes.

The glosses are primarily related to manual activity, not to body or facial activity, even though some lexical items include a specification of non-manual (mostly mouth) activities too. Non-lexical non-manual activity, as when the signer makes a manual sign accompanied by a head shake, are also not encoded by the gloss annotations: only the manual sign has been referred to in the gloss, not the negation expressed by the headshake.

3 Challenges

In order to improve machine search and machine processing of sign language corpora, several challenges need to be dealt with (Johnston, 2008). Most of these challenges stem from the fact that the glosses do not contain a transcription of the form of the language itself, but a pointer to a lemma in another language (Dutch, in this case).

The first challenge we have recently tackled in our NGT corpus concerns homonymy and polysemy. As for spoken languages, some signs have the same manual form, but do not share the meaning: homonyms, or do have the same manual form and have related but not identical meanings: polysemes. Lexicographers would define polysemes within a single dictionary lemma, while homonyms are treated in separate lemmata. In spoken English, the word 'arm' is an example of a homonym, which can refer to a limb, or it can be related to a weapon. In the first version of our sign language annotation conventions, homonyms and polysemes were ignored, as signs received a gloss based on the meaning of the manual part of the sign. In section 3, 'revising the annotation convention', we will discuss how homonyms and polysemes are currently processed. If homonym signs as well as polyseme signs would receive different glosses, an automatic recognition system would have severe difficulty grouping those signs that have the same forms.

A similar problem for recognition relates to the existence of sign variants: signs that have the same meaning, but a different form. Sometimes the same signers use these different signs as synonyms, but in addition there is some regional variation in the lexicon that the corpus recordings explicitly aimed to include. This type of variation was ignored in the initial release of the corpus as well, in that synonyms and/or regional variants simply received the same gloss.

Some additional challenges can be found in simultaneous constructions, whereby the left hand is articulating another sign than the right hand, which can even be one hand of a previous two handed sign (spreading) articulated simultaneously with a second sign. These types of special constructions are posing some real challenges for machine recognition and translation systems, as those constructions convey a large range of creative combinations, which cannot be translated easily, let alone consistently. Classifier constructions pose an additional serious challenge for machine processing. Classifiers are non-lexical signs, which refer to a category of referents and their location and/or motion, and they too can be translated in multiple ways. For example, the sign for car can be used at first, and when referring to the car later in the discourse when it is driving across a hill for example, NGT signers use a flat hand, moving up and down a virtual hill.

4 Revising the glossing conventions

4.1 General revisions

Based on the need to adapt the glosses to facilitate machine-readability, a series of revisions have taken place. A thorough check of typos and spelling mistakes has taken place. At the same time, minor revisions of the annotation conventions such as notating 'INDEX' as 'IX' were implemented. Secondly, the time alignments between the video and the glosses were checked and adapted where necessary.

4.2 Umbrella glosses

An important addition to the first version of the conventions concerns signs that have identical manual forms, but differ in mouth pattern. These form a very frequent group of manual homonyms and polysemes. Some signs can have multiple meanings, depending on the context and whether or not a mouthing is used (Schermer, 1990, Crasborn et al., 2008, van de Sande & Crasborn, 2009).

We refer to part of those identical sign forms with related meanings (polysemes) by adding what we call an 'umbrella gloss' to the more specific gloss (examples will be discussed below). Signs that have an identical manual form can thus be labeled with a more general name, while keeping the information about the meaning in context.

The advantage of this approach is that during the annotation process, we do not have to make decisions on the exact status of the combinations between manual and non-manual activities: whether or not they form independent lemmata with fixed meanings is left to further research, but we facilitate further research by including a reference to both the manual form (by the umbrella gloss) and the contextual meaning (typically invoked by the action of the mouth). In a sense, this approach forms a midway between using phonetic transcription and foreign language labels, as it represents both the unique identification of the form as well as reference to the contextual meaning of the sign.

As the process of annotation continues, the number of umbrella glosses will increase. 'AL' (ALREADY in English) is an example of such an umbrella gloss. The label 'AL' is being used for various signs with an identical manual form, but with (somewhat) different

meanings. To further specify the sign, an addition is used, for example AL:GEWEEST (AL:BEEN), AL:GEHAD (AL:HAD) or AL:AF (AL:FINISHED). As with any sign language gloss, an (umbrella) gloss is not a translation of a sign; it remains a label attached to the sign. In the absence of a complete and accessible lexicon, it facilitates consistency in the glossing. In fact, an umbrella gloss can be chosen rather arbitrarily, as long as the label is used consistently. Below, two examples are listed for such signs that belong to an umbrella gloss: ALREADY and PROGRAMME. On the left is the more neutral gloss (the umbrella gloss), on the right the glosses that can be used when for example an accompanying mouth pattern adds to the meaning of the sign. When the sign has no accompanying mouth pattern, the more neutral term (or umbrella gloss) is used.

Umbrella gloss	**Possible glosses if a sign has, for example, an accompanying mouthing.**
AL (ALREADY)	AL:GEHAD (HAD) AL:GEWEEST (BEEN) AL:AF (COMPLETED)
PROGRAMMA (PROGRAMME)	PROGRAMMA:REGELS (RULES) PROGRAMMA:WETTEN (LAWS) PROGRAMMA:EISEN (DEMANDS) PROGRAMMA:PLAN (PLAN) PROGRAMMA:AGENDA (AGENDA)

Figure 3. Simultaneous constructions involving numeral buoys

In the online NGT lexicon[4] these variants are not all listed as instances of a shared type; this is one of the reasons why it is hard to use an existing, fixed, lexicon for annotating a sign language corpus. The variation in the combinations of manual and non-manual forms can be used to further enhance existing lexicons.

4.3 True homonyms

A second part refers to identical forms as well, but instead of holding a shared meaning, these signs have highly distinct meanings: homonyms. An example from NGT is DOCTOR and BATTERY, which are both formed by the curved index and middle fingers touching the chin. Those types of homonyms will not share an umbrella gloss. For automatic sign recognition as well as for phonological and lexico-semantic research, it is crucial that such additional homonyms are listed separately as 'true homonyms', as separate from the polysemes that are joined by an umbrella gloss.

4.4 Regional variation in manual forms

Another addition to the conventions concerned sign translations that can have different sign forms, the so-called (regional and interpersonal) variants. It is important that different signs that have the same meaning (and therefore would receive the same gloss) but with a different sign form, can still be distinguished. The way to do this is adding a capital letter suffix to those glosses. A separate document is being made with different sign variants, for example: DOG-A and DOG-B.

4.5 Numeral constructions

Number signs were also in need of revised annotation conventions. Instead of glossing 'counting hand' for one hand, and 'IX' (pointing) for the other hand, we revised our gloss conventions so as to specify where exactly the dominant hand is pointing to. The gloss 'IX' is being used followed by a specification of the finger that is pointed at/indexed. Of this finger that is pointed at, only the first letter is glossed, e.g., IX:D (D for duim (in Dutch) = thumb) or IX:W (W for wijsvinger (in Dutch) = index finger). The non-dominant (counting) hand is specified for the number that is being realised, rather than merely stating 'counting hand'. See the example in Figure 3.

4.6 Spatial variability

Some lexical signs can be performed in highly distinct manner, for example for direction verbs, such as ASK and VISIT. The glosses were adapted, such that different glosses are given, depending on the direction of the verbs. If a sign is directed from the signer towards another addressee, the gloss is composed of 1GLOSS, for example 1ASK, whereas if a sign is directed from an addressee towards the signers, the gloss is composed of GLOSS1, for example, ASK1. The number 1 refers to

[4] http://www.gebarencentrum.nl

the signer.

4.7 Sentences

Sentence boundaries are clearly needed for machine recognition and translation. However, although a series of boundary cues were found in past research, final conclusions on boundary markers have not been established thus far (Crasborn 2007, Nicodemus 2009). In order to facilitate sign language recognition, sign language sentence boundaries based on intuitive judgments were added to the annotations. Moreover, translations were provided for one topic in the corpus and boundary cues are examined, specifically designed for a European project; SignSpeak.

5. Conclusion

As for all language corpora, sign language corpora should be created in a systematically controlled and consistent way, which make machine searches and machines processing possible (Johnston 2008). This not only provides us the possibility to study linguistic properties in sign languages into much more depth and using much larger sign data sets than before, but, importantly, it has also resulted already in first steps towards automatic sign recognition and sign to text translations. In order to achieve this, we have revised the glossing conventions of the first release of the Corpus NGT in such a way that they consistently label specific forms, taking into account creative variations of which it is not clear whether they have been lexicalised or not. In this way, we also try to circumvent the absence of a well-accessible lexicon.

It will be clear from the discussion in this paper that we have aimed to create a workable solution that addresses the demands of both linguistic research and language technology. Further discussion on both details and principled choices is clearly necessary. A workshop of the *Sign Linguistics Corpora Network* in June 2010 is devoted to annotation, and will also take up the discussion on sign language glossing.[5]

5 References

Crasborn, O., J. Mesch, D. Waters, A. Nonhebel, E. van der Kooij, B. Woll & B. Bergman (2007) Sharing sign language data online: experiences from the ECHO project. *International Journal of Corpus Linguistics* 12:535-562.

O. Crasborn & T. Bloem (2009) Linguistic variation as a challenge for sign language interpreters and sign language interpreter education in the Netherlands. In Jemina Napier (ed.) *International perspectives on sign language interpreter education* (pp. 77-95). Washington DC: Gallaudet University Press.

Crasborn, O. & H. Sloetjes (2008) Enhanced ELAN functionality for sign language corpora. In: O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd & I. Zwitserlood, eds. *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*. Paris: ELRA. Pp. 39-43.

Crasborn, O. & M. de Wit. 2005. Ethical implications of language standardisation for sign language interpreters. In *International perspectives on interpreting. Selected proceedings from the Supporting Deaf People online conferences 2001-2005*, edited by J. Mole. Brassinton: Direct Learn Services, pp. 112-119.

Crasborn, O. & I. Zwitserlood (2008) The Corpus NGT: an online corpus for professionals and laymen, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd, eds. ELDA, Paris, pp 44-49.

Crasborn, O., I. Zwitserlood & J. Ros (2008) Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands. Centre for Language Studies, Radboud University Nijmegen. http://www.ru.nl/corpusngtuk

Hanke, T. (2002). iLex - A tool for sign language lexicography and corpus analysis. In: M. González Rodriguez, Manuel and C. Paz Suarez Araujo (eds.): Proceedings of the third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain; Vol. III. Paris: ELRA. 923-926.

Hanke, T., Konrad, R., & A. Schwarz, S. (2001). GlossLexer – A multimedia lexical database for sign language dictionary compilation. Sign Language and Linguistics 4(1/2): 161–179.

Johnston, T. (2008) Corpus linguistics and signed languages: No lemmata, no corpus. In: Crasborn, O., Hanke, T., Thoutenhoofd, E.D., Zwitserlood, I. and Efthimiou, E. eds. *Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the representation and processing of sign languages*. Paris: ELRA, pp. 82-87.

Johnston, T. (2009) Guidelines for annotation of the video data in the Auslan Corpus. Ms., Macquairie University. http://media.auslan.org.au/media/upload/attachments/Annotation_Guidelines_Auslan_CorpusT5.pdf.

Johnston, T., Vermeerbergen, M., Schembri, A., & Leeson, L. (2007). 'Real Data Are Messy': Considering Cross-Linguistic Analysis of Constituent Ordering in Australian Sign Language (Auslan), Vlaamse Gebarentaal (VGT), and Irish Sign Language (ISL). In P. Perniss, R. Pfau & M. Steinbach (Eds.), *Proceedings of the Workshop on Sign Languages: A Cross-Linguistic Perspective, Mainz, Germany, March 25-27, 2004*. Berlin: Mouton de Gruyter.

Miller, C. (2001) Some reflections on the need for a common sign notation. *Sign Language & Linguistics* 4:11-28.

Sande, I. van de & O. Crasborn (2009) Lexically bound mouth actions in Sign Language of the Netherlands. A comparison between different registers and age groups. *Linguistics in the Netherlands* 26: 78-90.

[5] http://www.ru.nl/slcn

Schembri, A. (2008) British Sign Language Corpus Project: Open access archives and the observer's paradox, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitserlood & E. Thoutenhoofd, eds. ELDA, Paris, pp. 165-169.
Schembri, A. & O. Crasborn (this volume).
Schermer, Trude. 2003. From variant to standard. An overview of the standardization process of the lexicon of Sign Language of the Netherlands (SLN) over two decades. *Sign Language Studies* 3 (4): 96-113.