From corpus to lexicon: the creation of ID-glosses for the Corpus NGT

Onno Crasborn, Anne de Meijer
Centre for Language Studies, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl, a.demeijer@let.ru.nl

Abstract

When glossing of the Corpus NGT started in 2007, there was no lexicon at our disposal to base ID-glosses on. Semantic labels were used without ensuring a constant relationship between sign form and gloss. This is currently being repaired by creating a lexicon from scratch alongside with the creation of new annotations. This substantial task is still in progress, but promises to lead to several new research avenues for the future. The current paper describes some of the choices that were made in the process, and specifies some of the glossing conventions that were used.

Keywords: sign language, corpus annotation, gloss, ID-glossing, lexicon, lexical database

1. Introduction

In the first release of the Corpus NGT in 2008, a set of 64,000 glosses for 163 sessions was included in the online Language Archive at the Max Planck Institute for Psycholinguistics.[1] Like the media files, the annotation files for most sessions have become publicly accessible. Providing an EAF file for every session, also the non-annotated ones, enables users to view the synchronised movies and any available annotations in their browser, using the ANNEX tool.[2] ANNEX allows for similar searches as ELAN, both in single files and within and across corpora.

As the glosses were created by a diverse group of mostly linguistically naïve signers that were insufficiently supervised and monitored, the resulting annotations were of variable quality. Moreover, for many aspects of the glossing, insufficient explicit guidelines were available. This paper describes the various steps that were taken to improve the glossing since then, including the present glossing conventions, working towards a second release of a larger set of annotation documents with ID-glosses later in 2012.

2. General issues in glossing signed interaction

Unlike the documentation of spoken languages, samples of sign language utterances are typically not glossed in the language itself, lacking a commonly used writing system or phonetic notation system. Occasionally, researchers have used HamNoSys for this purpose. More commonly, researchers create glosses in the writing system of a spoken language, whether it is the spoken language known to the deaf community in question or the language of the publication, typically being English. The choice of the spoken language word is typically not crucial, as long as it is a label that is semantically interpretable with respect to the sign by the target audience. By consequence, it may be variable across publications, and moreover it is uninformative about the form, the precise meaning, or the function of the signed word in that particular context.

While such a strategy is efficient for presenting example sign sentences in text documents, Johnston (2008) argues that it would be unwise to go about in such a loose way when annotating sign corpora. More precisely, it is key that all instances of the same sign lemma or the same full form are represented by the same word. In fact it does not matter what this word is, and it could well be a unique number. As long as a unique identifier is used, the form in context can be related to a lexicon. Johnston calls such glosses 'ID-glosses'. They primarily serve the purpose of providing a unique written identifier for every sign instance. In practice, both for annotation of new texts and for the interpretation of annotated texts, written words form the most practical solution for the identification problem, even though they may provide the false impression that the full semantics of a form in context is covered by the written word that forms the gloss.

Crasborn, Hulsbosch & Sloetjes (2012) describe a technical solution within the ELAN annotation software that in fact does use a numerical identifier for every gloss; it is this identifier that links a given annotation on a gloss tier to an external XML file that contains a list of lexical items. The surface form of this unique identifier that users see is still a text string. This in fact works not only for glosses, but for any tier in an annotation document for which a (external) controlled vocabulary can be defined.

3. Glossing of the Corpus NGT

3.1 Initial procedure

While it is clear that one cannot get around using ID-glosses in creating a machine-readable linguistic

[1] As the left and the right hand are both assigned a gloss annotation in the case of two-handed lexical items, the total estimated number of signs is 49,000, of which 15,000 are two-handed and 34,000 are one-handed.

[2] ANNEX can be opened from the corpus browser at http://corpus1.mpi.nl: in the contextual menu of an annotation document, an option appears to view the node or to perform an annotation content search.

corpus, actually using them for resources of a particular language is contingent on the existence of a lexicon that one can refer to with unique identifiers. As no lexicon was readily available to the annotators of the Corpus NGT to select glosses from and to add new glosses to, the original task for annotators was to create a translation of the form in context that appeared to be most fit to the core meaning of the sign. Thus, specific contextual meanings of the sign should not appear in the glosses. While it was recognised at the time (2007-2008) that some variation in the selection of glosses of any given form would ensue, our hope was that it would be relatively easy to take into account such variation when the corpus would be used for research later on. For example, when searching for a sign with a specific phonological form, the researcher would always be aware that different glosses for that form would have been used, and adapt his searching strategies accordingly. While there may be some value in this approach, it still requires a substantial amount of interpretation and action from researchers. We gradually acknowledged that this would never lead to a truly machine-readable corpus for the lexical level. As the signed word is such a basic unit that will be involved in nearly any linguistic or technological study, machine-readability is especially crucial at this level. We therefore decided to create a lexicon specifically for the Corpus NGT annotations.

3.2 A lexicon for NGT corpus annotations

The creation of the Auslan corpus (Johnston 2008b) started long after lexical resources for this language were developed by the same researcher (Johnston 1998, 2001). Thus, not only was there systematic knowledge of the Auslan lexicon, there was also a published resource from the same team that could form the basis for ID-glosses. For NGT, there is no open access reference lexicon. The existing lexical resources published by the Dutch Sign Centre are not available for research purposes, nor were they created as such. Different subsets have different origins, often created for educational purposes. The glosses that are used are targeted at easy use by laymen in a computer interface or paper dictionary, rather than at efficient computer processing. In addition, it is not unlikely that the selection of signs does not cover the lexicon that is used in the recordings of the Corpus NGT. Most crucially, this lexicon could not be expanded during the process of corpus annotations, simply because the workflow of the Dutch Sign Centre is quite different from that of the annotation of the Corpus NGT. For these various reasons, it was decided to start to compile a lexicon specifically for the Corpus NGT.

The lexicon started as a simple Excel sheet compiling ID-glosses for (regional or other) variants and a rough phonological description of each of them. This is currently being expanded to include all glosses, including semantic categories that do not have variant forms (see 3.3.1 below). To facilitate the selection of the correct gloss for a particular sign form, three fields were added. The first one contains other possible Dutch translations of the same sign form. A second column displays NGT homonyms, to point out that the same sign form has multiple glosses for distinct meanings of the sign. A third column contains related ID-glosses (by form or meaning) that may easily be confused with one another because of resemblance in form, meaning, and/or function; this information is especially useful for creating new annotations.

The added value of a corpus-based lexicon like this one is reflected by the column with Dutch translation variants. Information in this column is not just composed by making up possible Dutch translation variants of the gloss, but also contains translations actually used for that sign form, originally by annotators in the phase of intuitive glossing and currently by annotators who create annotations on the child tier 'Meaning' for a gloss. At this moment, we have not yet developed an automatic way of harvesting these meanings specified for glosses.

Currently, a phonological description has been created for every ID-gloss. The translation variants, homonym, and related glosses columns are used extensively. Further, multiple other columns for additional information are created. Whenever information is available, we specify the origin of a sign (a specific region in the Netherlands, derived from fingerspelling, a gesture, an ASL loan, etc.), the image a sign depicts can be described (COFFEE displays the image of grinding of coffee beans), mouthings or mouth gestures can be added, and observed or known phonetic-phonological variation can be specified, such as one-handed occurrences of a sign described as two-handed. As the lexicon gradually grows, we expect the use of these columns to also increase.

This Excel-based lexicon is soon to be converted to the lexical database LEXUS, probably with more structure and built-in links to related glosses. A video clip of a citation form of each entry will need to be added, as well as links to instances of the full form in context in the Corpus NGT. An area of concern is the ease of updating the lexicon once it is in LEXUS; this will no doubt be less efficient than in Excel.

3.3 Additional annotation conventions for glosses

In the following paragraphs, we briefly characterise the various annotation conventions pertaining to gloss annotations. They will appear in a more detailed form with further description in Crasborn & de Meijer (in prep.).

3.3.1 General form and labelling of variants

The general form of glosses is a single Dutch word written in capital letters. The word used for the gloss is the most neutral choice with respect to meaning and grammatical marking. To distinguish between signs with the same meaning, but different forms, alphabetical suffixes are used. For example, there are entries for HOND-A, HOND-B, and HOND-C, being three

different signs that all mean 'dog'. Signs with the same form, but unrelated meanings (homonyms) each receive their own gloss.

3.3.2 Signs vs. gestures

The lexical or gestural status of some sign forms is not easily determined. We consider gestures to be communicative hand movements that either are also used by the community of hearing non-signing speakers of Dutch, and/or that do not have a form-meaning relationship that can be described, such as beat gestures. Emblematic gestures that can be lemmatised (i.e., that have a root form and a meaning or other communicative function that can be listed) are treated like lexical items and are marked in the lexicon as (possible) gestures. All other potential gestures are marked by a percentage character (%) on the gloss tier. In this way, (possible) gestures can easily be retrieved and inspected more closely should this be relevant to an investigation; alternatively, they can be left out in automatic processing of corpus data altogether.

3.3.3 Morphologically complex forms

Morphologically complex forms like classifier constructions or depicting signs cannot be annotated using an ID-gloss, due to their highly context dependent form and meaning. However, at least some of their components do have a constant form-meaning relationship that can be described. Classifiers are glossed by a three-partite combination of 1) movement, 2) type, and 3) handshape. Thus, the annotation consists of three consecutive codes, separated by an underscore '_'. MOVE_EC_1 for example is a classifier moving through space, representing an entity, that has an extended index finger as its handshape. It thus likely refers to a long and thin entity moving through space, possibly a person.

Each combination is listed in the lexicon, to facilitate data entry and avoid typos. Although some aspects of the form are described by the gloss, the meaning is left unspecified in the lexicon: there are no translation variants of the combinations. Therefore, for these glosses, the child tier dedicated to meaning always needs to be filled in, with a compact description. In the example above, this could be 'person moves forward', for instance. Signed constructions whose handshape and movement show the specific shape of a referent ('size and shape specifiers') are glossed in a similar manner.

3.3.4 The Modification tiers

Further modifications of the movement or other components of the constructions that we just discussed can equally be characterised on a child tier. For every gloss tier (one per hand), there is a 'Modification' tier that allows for a textual description of the modification. These tiers can be used for all types of signs, not just the morphologically complex ones. If the example form in section 3.3.3 would be modified by an arced movement expressing 'jumping forward', for instance, this would be encoded here, rather than by altering the MOVE component. The latter serves to distinguish movement through space from being at a specific location (AT), mere presence (BE), and action without a path movement (ACTION).

At present, we do not yet have specific annotation guidelines for the Modification tiers. We recognise that this would be beneficial at some point, distinguishing systematic recurrent modifications with a clear describable form from more idiosyncratic pantomimic modifications of (parts of) signs. Our strategy is to first let people intuitively use the tiers, and then after some time investigate what type of distinctions are created.

3.3.5 Some further conventions

Just as it is most practical to use words instead of unique numbers as glosses, for some categories of signs it is practical to use additional conventions regarding glossing. Although every unique form receives its own gloss, the conventions group together certain lexical or morphological categories to facilitate retrieval.

Examples of such further conventions:

- Hyphens (-) are used to separate multiple words representing a single gloss, whereas underscores (_) are used in glossing morphological complex forms (see section 3.3.3 above).
- Pointing signs start with the basic gloss PT; several types of pointing signs are specified in the lexicon.
- Compounds of two or more sequential parts are glossed by separate annotations for each part, and are marked on the meaning tier using '^'.
- Lexical negation is marked using the suffix -NIET 'not', so that the regular and negated forms are next to each other in various alphabetically sorted lists, such as in search results or in sorted presentations of the lexicon.
- Numbers are always glossed using digits.
- Name signs are preceded by an asterisk (*).
- Fingerspelling is marked by a hedge mark (#).
- Uncertainty on the part of the annotator is marked by a question mark (?) following the gloss; such glosses do not receive an ECV link (see Crasborn, Hulsbosch & Sloetjes, this volume).
- Double question marks (??) are used for unknown signs. These annotations should periodically be inspected by native signers other than the annotators in order to determine their nature.

3.3.6 A comparison with the Auslan annotation guidelines

As made explicit by Schembri & Crasborn (2010), it is desirable to work towards some kind of standardisation of annotation conventions for sign language corpora, in order to facilitate cross-linguistic research and to promote the use of published resources by other research groups. We have attempted to copy many of the published conventions for the Auslan corpus (Johnston,

2011).

Comparing the two sets of guidelines, the major correspondence is in the annotation of the basic gloss, based on ID-glosses linked to a lexicon. Other conventions, like listed in 3.3.4, show some minor differences relating to how the information is encoded. Whereas the annotations in the Corpus NGT mainly group together certain categories by using a single generic character, in the Auslan Corpus this information is mostly coded by additional information separated from the gloss by a colon or put between brackets. However, overall, the same type of information is annotated. We choose to relegate additional information about a sign to dependent tiers as much as possible (including Meaning, Modification, Handshape, and Location), so that there can be a link to the External Controlled Vocabulary for each form. We hope that separating the ID-gloss from additional information will facilitate automated processing of annotations, whether within ELAN or by creating scripts that work directly with the EAF files.

3.4 Discussion: conflicting principles

Ease of processing is thus an important consideration in these glossing conventions, complementing linguistic principles and sometimes conflicting with them. This is true for the very essence of the ID-gloss, a (combination of) spoken language word(s) that may not be semantically identical to the sign, but it also holds for the glossing of compounds, for instance. Our current lexicon assumes that every entry consists of a single sign syllable. For every syllable, all phonological features can be described in a uniform way, avoiding the complexity of multiple syllables that have different hand configuration or location properties, for instance. As NGT has very few compounds or other signs consisting of sequences different syllables (van der Kooij & Crasborn, 2008), there are not many signs for which the workaround for annotating compounds is problematic. But for all those that do exist, the properties of the component parts can be more easily processed. When calculating frequencies of handshapes, for instance, the handshapes of single glosses are now automatically taken into account, regardless of whether the handshape is part of a compound or not.

4. Applying ID-glosses to an already annotated corpus

In several rounds of revision, we are currently building the lexicon list that corresponds to the signs used in the Corpus NGT, agreeing upon the ID-glosses at the same time. We have gone through many stages in this tedious process, from spell-checking to the creation of specific conventions for name signs, for instance. The current situation of early 2012 is that about 80% of the more than 120,000 gloss annotations has a reference to the lexicon. These include the most frequent signs in the sessions that are glossed until now, as it is these that we started to assign ID-glosses to first. The remaining 20% consists of various categories, representing both known and unknown variation and known and unknown errors. Glosses referring to complex signs (depicting signs, modified lexical signs, pantomime) number about 10,000; their annotation as described in section 3.3.3 will still take quite some time. A much smaller set consists of signs that were unknown to the annotators at the time of first annotation, and are marked by double question marks; these will need to be inspected by one or more native signers. The largest proportion however, an estimated 20,000 glosses, are expected to consist of signs that are simply used infrequently. There may still occasionally be singleton glosses that refer to existing items in the lexicon, but we expect that most of them will be infrequent signs that have yet to be added to the lexicon. The lexicon currently counts 1,800 items; we expect it to grow to 4,500 by the time that these infrequent signs have all been inspected.

The lowest level of corrections, repairing typos and spelling mistakes, is slowly becoming less necessary now that a controlled vocabulary is used for the gloss tiers, making manually typed input less and less necessary (see Crasborn, Hulsbosch & Sloetjes, this volume).

A challenge that does persist and that is beyond the current round of establishing ID-glosses, is deciding whether all the variants and homonyms that have been created in the lexicon are actually independent lemmata or not. As we indicated above, some decisions on when to create a new lexical item were made to facilitate automated processing, but in other cases there was simply a lack of knowledge about (the uses or meanings of) a lexical item. It is here that the most difficult task lies as soon as a certain level of consistency is guaranteed, and it is a challenge for present users of the corpus to take the nature of the existing lexicon into account. We see this as a consequence of developing a lexicon and a corpus in tandem, and improving the nature of the lexicon will remain a rocky road for some years to come.

5. Expected developments

5.1 A second public release of the Corpus NGT annotation files

A second release of the Corpus NGT annotations will be made public in the Language Archive as soon as the unknown territory of 20,000 glosses has been inspected, hopefully before the end of 2012. The aim for this second release is not to definitively establish ID-glosses for all signs, but rather to make explicit which glosses still need closer inspection and should thus be treated with caution. This will include signs that have not been identified, but also complex constructions that have yet to be described in terms of component parts in the way outlined in section 3.3.3.

Upon that second release, the accompanying lexicon will be published in the online tool LEXUS as well as in the form of an external controlled vocabulary on a web server, linked to the gloss tiers. Moreover, the

broader annotation conventions, including those for glosses summarised in this paper will be published in the form of a larger document (Crasborn & de Meijer, in prep.).

5.2 Development of lexicons

The RU lexicon will also be further enlarged during the annotation of other resources than the Corpus NGT, including an on-going data collection of longitudinal recordings of deaf parents with their children.

In the context of future research projects, we further hope to explore the option of the integration of lexicons of different signed languages within LEXUS. Moreover, we hope to create an English version of all the ID-glosses, and explore ways of switching between languages in ELAN for annotations like ID-glosses that should ideally be multilingual, much like ISOcat data categories may have multiple language sections. Generating ISOcat data categories for lexical items might in fact be a strategy to address this wish, and it may also facilitate multilingual lexica in the sense of a 'universal sign dictionary' ('universal SignBank', Trevor Johnston, pers. comm.): there could be a data category for a specific form that can have different meanings or functions in different languages.

6. Acknowledgements

The work described in this paper has been supported by ERC Starting Grant no. 210373 and NWO VIDI grant no. 276-70-012 awarded to Onno Crasborn, by EU grant 231424 'SignSpeak', and by the Centre for Language Studies of Radboud University Nijmegen. We thank Micha Hulsbosch for his help in automating tasks related to revising glosses. In alphabetical order, we thank our colleagues Richard Bank, Lianne van Dijken, Els van der Kooij, Yassine Nauta, Ellen Ormel, Johan Ros, Anna Sáfár, Merel van Zuilen, and Inge Zwitserlood for their contributions to the discussions on glossing conventions and the workflows that we followed. Moreover, we thank Karlijn Hermans for her substantial contribution to the correction of errors and the tagging of variants in the corpus.

7. References

Crasborn, O., Hulsbosch, M., & Sloetjes, H. (2012). Linking Corpus NGT annotations to a lexical database using open source tools ELAN and LEXUS. Paper presented at the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon.

Crasborn, O. & de Meijer, A. (in prep.) Annotation conventions for the Corpus NGT. Ms., Radboud University Nijmegen.

Johnston, T. (1998). Signs of Australia. A new dictionary of Auslan (Australian Sign Language) (the sign language of the Australian deaf community). North Rocks, NSW, Australia: North Rocks Press, Royal Institute for Deaf and Blind Children.

Johnston, T. (2001). The lexical database of Auslan (Australian sign language). Sign Language & Linguistics, 4(1/2), 145-170.

Johnston, T. (2008a). Corpus linguistics and signed languages: no lemmata, no corpus. In O. Crasborn, E. Efthimiou, T. Hanke, E. Thoutenhoofd, D. & I. Zwitserlood (Eds.), 5th Workshop on the Representation and Processing of Signed Languages: Construction and Exploitation of Sign Language Corpora (pp. 82-87). Paris: ELRA.

Johnston, T. 2008b. Corpus of grammar and discourse strategies of deaf native users of Auslan (Australian Sign Language), Endangered Languages Archive, SOAS, University of London. http://elar.soas.ac.uk/deposit/johnston2012auslan

Johnston, T. (2011). Auslan Corpus Annotation Guidelines. Unpublished manuscript, Macquairie University, Sydney. http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf

Kooij, E. van der, & Crasborn, O. (2008). Syllables and the word prosodic system in Sign Language of the Netherlands. Lingua, 118, 1307-1327.