A corpus-based lexical database for Sign Language of the Netherlands

Onno Crasborn, Inge Zwitserlood, Els van der Kooij, & Richard Bank
Radboud University, Centre for Language Studies, PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl, i.zwitserlood@let.ru.nl, e.van.der.kooij@let.ru.nl, r.bank@let.ru.nl

Abstract

This paper describes the creation of a corpus-based lexical database for a signed language, namely Sign Language of the Netherlands, which can later be developed into a multipurpose dictionary. The content of the database is based on the annotation of a large video corpus of this language, using ID-Glosses as the key between database and annotation files by lack of an accepted writing system. Our practices contribute to developing international conventions for both glossing practices and lemmatisation of signed languages, while being innovative in the way frequency data from the corpus are included in the lexical database. The paper further describes some of our recent contributions in the development of the Signbank software and the ELAN multimedia annotation tool.

Keywords: corpus-based lexical database, sign language, sign language lexicography, Sign Language of the Netherlands (NGT)

1. Introduction: Sign Language Corpora and Corpus-Based Lexical Resources for Sign Languages

Sign language dictionaries have been produced for many different sign languages in the last five decades, initially only in paper format with drawings or photos (see Carmel (1992, 1994), Zwitserlood (2010), and Zwitserlood et al. (2013) for overviews), but since the 1990s also in various electronic formats (e.g. Savolainen, 1995; Brien et al. 1995; Johnston, 1997; Hanke, Konrad & Schwarz, 2001). As Vale (2010) and Zwitserlood (2010) observe, such dictionaries are typically bilingual (signed language–spoken language), even though the titles often suggest they are monolingual. The target audience of the dictionaries is often primarily hearing second language learners. There are many such L2 learners of most sign languages, if only because most deaf children are born to hearing parents who start learning the local sign language early in their child’s life. The bilingual nature of the dictionaries together with the lack of a commonly used writing system has often (but not always) meant that dictionaries mostly used spoken language for sorting and searching. Through lack of natural
language sources these dictionaries have often been based on elicitation studies, sometimes specifically targeted at establishing the amount of (regional) variation (e.g. Stroombergen & Schermer, 1985, for Sign Language of the Netherlands, henceforth NGT, Nederlandse Gebarentaal). Corpus-based lexicography has begun in the last decade. Corpora based on video data have only started to become a reality since around 2000 (Crasborn 2008). Currently, there are a number of corpora published online, and more under development. Because sign languages do not have accepted writing systems, these corpora are typically annotated with use of ‘ID-glosses’, i.e. labels consisting of words of a spoken language that approximate (one of) the meaning(s) of the signs. Such a technique necessitates the creation and use of a lexical database that is linked to the corpus annotations (Johnston, 2008, 2010). The primary goal of such lexical databases is often to enable corpus annotation, but they may develop into a sophisticated bases for dictionaries. This paper will describe the creation of a corpus-based lexical database for NGT, called NGT Signbank.

2. NGT Signbank

The NGT Signbank web application is an offspring of the Auslan Signbank software (see http://github.com/signbank for these and related projects). It is located at http://signbank.science.ru.nl, and is at present accessible only to researchers. The aim has been to create a lexical resource to annotate and link to NGT corpora, including the Corpus NGT (Crasborn, Zwislerood & Ros, 2008; Crasborn & Zwitserlood, 2008). The third release of the annotation for this largest digital resource for NGT was published in June 2015, containing approximately 90,000 sign tokens (Corpus NGT Researchers Site). The corpus annotations (Crasborn et al., 2015)) first and foremost serve a scientific purpose, facilitating linguistic and technological research on NGT. As the Corpus NGT is an open access collection of texts, accessible through The Language Archive (https://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6@view), it is also accessible for creating applied linguistic resources such as dictionaries. At present, the NGT Signbank database contains over 3,900 lexical entries with rich phonological descriptions, some 500 of which have yet to be discussed (see below). The system has a trilingual interface, in English, Dutch, and Mandarin Chinese. Software development is on-going, and future versions will include a morphology component and the possibility of using multiple datasets of different languages in a single database. Plans are being developed for merging the different forks of the original Auslan Signbank code back into a single code-base, creating a single configurable system that could be set up for any signed language.

3. Towards a Corpus-Based lexical database for NGT

3.1 Building up a database from corpus instantiations
Lexical entries in Signbank are added step by step as corpus annotation proceeds: lexical items that are not yet present in Signbank are assigned an ID-gloss proposal by the annotator, enriched with a temporary movie and translation equivalents in Dutch. Regular meetings between (at least) a linguist and two deaf annotators discuss new proposals, double-checking whether the sign is truly lexicalised, whether no overlaps with existing entries occur, and whether gloss and translations are appropriate. To a large extent, lemmatisation is performed in this stage, even though the ID-glosses used for annotations are not targeting the lemma level but the full-form level. A Lemma ID-Gloss field is available for lemmatisation beyond this stage, grouping the various pointing forms under a single Lemma ID-Gloss ‘PT’, for instance.

3.2 Translation equivalents

Translation equivalents are manually selected based on the intuitions of deaf native signers who are responsible for the annotation, and who encounter a sign in context. Upon discussion about new entries in the database, these translations are evaluated and extended where possible. Although gloss annotations in the corpus can have the property ‘meaning’ (a child annotation containing the meaning in context), these contextual meanings are not automatically harvested to enrich the database: manual selection is important to highlight the core meanings of a sign while disregarding contextual collocations of manual signs and independent mouthings (see below). In the future, these translations could be complemented by a field that does indeed show these observed meanings, including their frequencies.

3.3 Mouthings

The ID-glosses in the Corpus NGT and thus the entries in NGT Signbank refer solely to manual signs. Articulations of spoken Dutch (mouthings) that accompany signs in the corpus are considered to be words from that spoken language. Recent research on the Corpus NGT data showed that while there is considerable consistency in the selection of the word for some signs, there is substantial variation for others (Bank 2015; cf. Schermer 1990). All mouth actions are annotated on separate tiers in the corpus. Whereas in sign language dictionaries citation forms are usually presented with an articulation of the typical or most frequent spoken word, or one that best matches the most general meaning of the sign, citation form movies in NGT Signbank do not show articulations of mouthings, emphasising the focus on the manual nature of the lexical entry (see also Schermer (2001) for further discussion). It turns out to be hard during corpus annotation to distinguish the meaning of the manual sign and the semantic contribution of the mouthing, and many contextual meanings linked to glosses are in fact the composite meaning of the manual sign and the mouthing.

3.4 Frequency of use
The frequency of use of gloss annotations is automatically harvested from the corpus every night, generating several statistics. Token frequencies are calculated over the whole corpus and for each of the regional variants as distinguished in the metadata. These are the five traditional dialect regions in the Netherlands (Schermer, 2004), and a sixth category that includes signers with a mixed regional profile. The number of signers that produce tokens of a sign is also calculated, for the whole corpus and per region. This second type of information is particularly useful in determining how widespread the use of a sign is within a region, distinguishing idiosyncratic uses by a single signer from systematic use in the specific region. For further discussion, see Crasborn et al. (2016).

4. Interaction between Signbank and ELAN

Although initial steps to display the content of Signbank while working in ELAN have already been taken in earlier releases of ELAN (Crasborn, Hulsbosch, & Sloetjes, 2012), on-going software development is aimed at establishing a more direct link between the offline annotation tool ELAN and the web service Signbank. Using the existing Lexicon Services function in ELAN, users will be able to specify a Signbank location, upon which a variety of functions becomes available. All of these are aimed at improving the consistency of the annotation process and facilitating access to the lexical database from annotated files. A new ELAN release with this functionality is foreseen for the first half of 2017. On the Signbank end, database content needs to be made accessible to this web service. This is presently in a testing phase for NGT Signbank and ASL Signbank, the latter still being under development.

5. Conclusion

The above developments illustrate how the availability of sign language corpora is gradually leading to corpus-based lexical databases. It is to be expected that these in turn will lead to corpus-based dictionaries that are also linked to these corpora. This will allow for new types of learning and teaching materials for L2 users, but hopefully also for a type of dictionary that is of interest to deaf people themselves. In addition, a lexical database of the kind described here together with an annotated corpus can form the basis for further language resources.

6. References


Crasborn, O., Zwitserlood, I., & Ros, J. (2008). The Corpus NGT. An open access digital corpus of movies with annotations of Sign Language of the Netherlands (Video corpus). from Centre for Language Studies, Radboud University Nijmegen http://hdl.handle.net/1839/00-0000-0000-0004-DF8E-6


This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

[http://creativecommons.org/licenses/by-sa/4.0/]