NGT Signbank

Data · January 2015

DOI: 10.13140/RG.2.1.2839.1446

12 authors, including:

Onno Crasborn
Radboud University
143 PUBLICATIONS **620** CITATIONS
SEE PROFILE

Els van der kooij
Radboud University
31 PUBLICATIONS **300** CITATIONS
SEE PROFILE

Ellen Ormel
Radboud University
33 PUBLICATIONS **297** CITATIONS
SEE PROFILE

Linking lexical and corpus data for sign languages: NGT Signbank and the Corpus NGT

Onno Crasborn, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anique Schüller, Ellen Ormel, Ellen Nauta, Merel van Zuilen, Frouke van Winsum & Johan Ros

Radboud University, Centre for Language Studies
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands

Keywords: sign language resources, multimodal annotation, Signbank, ELAN annotation software

1. Introduction

How can lexical resources for sign languages be integrated with annotated video corpora? In this paper we aim to answer this question by discussing an increasingly frequent scenario for sign language resources, where the lexical data are stored in an online lexical database, while the annotation data are offline files in the ELAN Annotation Format (EAF).

Lexical databases for sign languages often originated from the purpose of creating sign language dictionaries (Johnston, 2001). These dictionaries were created in a variety of contexts, ranging from language technology or linguistics departments within academia to deaf associations. The varying demands and facilities have led to a diversity of proprietary databases and some open source solutions. A standard even for data structures in this domain is not within view. It is therefore important to document existing solutions, as we do in this paper.

In terms of annotating and retrieving lexical signs for linguistic research, there is by now broad consensus on the need for ID-glosses (Johnston, 2008, 2010) in corpus annotation, which in turn requires having at least a list of ID-glosses with a description of the phonological form and meaning of the signs.

This paper contributes to the establishment of standards for sign language resources by discussing how two data resources for Sign Language of the Netherlands (Nederlandse Gebarentaal; NGT) are currently being integrated, using the ELAN annotation software for corpus annotation (Wittenburg et al., 2006) and an adaptation of the Auslan Signbank[1] software (Johnston, 2001, 2010) as a lexical database.

2. Two existing data sets

This section describes first the Corpus NGT and then NGT Signbank. While not the only option, as we will discuss in the conclusion, this type of combination of data sets is getting more common in the domain of sign language resources.

2.1 Corpus NGT

The Corpus NGT (Crasborn & Zwitserlood, 2008; Crasborn, Zwitserlood, & Ros, 2008) is a collection of video and annotation data of 92 prelingually deaf signers, recorded in dyads, who retell video clips and picture stories and discuss issues related to deafness, deaf education and sign language. Annotation of the corpus is on-going; the latest (third) public release of Corpus NGT annotations that was published in June 2015 (Crasborn et al., 2015) contains over 145,000 glosses for the left and right hands.

2.2 NGT Signbank

The Signbank lexical database software has originally been developed for Australian Sign Language[2] (Auslan; Johnston, 2001), and has since also been implemented for British Sign Language[3] (BSL; Fenlon et al., 2015) and NGT[4] (Crasborn et al., 2014).

3. Existing relations between data sets

This section describes two types of relationships between corpus and lexical database that are already implemented, while section four will focus on some further interactions between the data sets that will make exploitation of the data richer and easier. All interactions are visualised in Figure 1.

3.1 The lexical database as a vocabulary of gloss types for annotation

To facilitate video annotation in ELAN, an external controlled vocabulary (ECV) is used. An ECV contains the full list of ID-glosses in Signbank, to label lexical signs with, as well as phonological information about those signs (e.g. handshapes, location, movement direction) and Dutch translation equivalents that serve to clarify their meaning. When deciding on an annotation, the annotator chooses an entry from the ECV to be included in the EAF file. This facilitates decision-making and reduces the occurrence of typing errors. The ECV is centrally stored on a web server (hence the E for external), allowing for central updating of the ECV with changed or added glosses, phonological information and meaning. The ECV is automatically reloaded each time an EAF file is opened on a local computer with an internet connection. Annotation values of the glosses are then updated, if applicable, to reflect the current information in the ECV. This ensures that the annotators always work with the latest version.

[1] http://www.auslan.org.au

[2] https://bitbucket.org/stevecassidy/signbank/

[3] http://bslsignbank.ucl.ac.uk

[4] http://signbank.science.ru.nl

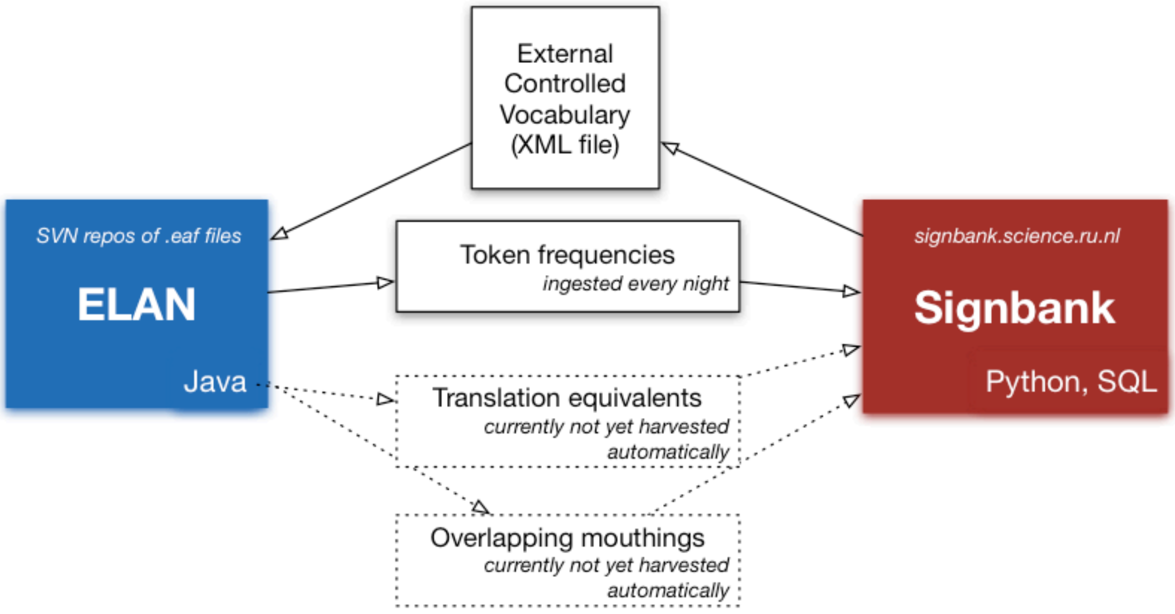

Figure 1: Overview of existing and foreseen relations between ELAN and Signbank.

A bilingual ECV is generated directly from Signbank. A nightly server-side script generates an updated ECV, including all changes made to the Signbank database in the previous day.

The relation between Signbank and EAF files through the ECV is currently one-way: items are added to Signbank and then displayed in ELAN. It is not yet possible to harvest new items in ELAN files and add these to Signbank, for instance, or to manually add a new item to Signbank from within the ELAN interface. Both of these options will be explored in the near future.

3.2 Token frequencies in the lexicon

Two types of frequency data are automatically ingested in Signbank from the glosses in the Corpus NGT. First of all, there are token frequencies over the whole corpus and for each of the six regions distinguished in the metadata. Second, the number of signers that produce tokens of a sign is also calculated and ingested in Signbank, 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: is it an idiosyncratic (perhaps older) form used by a single signer, or are there several people using the same sign?

4. Data set interactions to develop

We presently foresee three types of data interactions that could be implemented fairly easily, and that would enrich Signbank on the basis of corpora.

4.1 Harvesting of translation equivalents in the corpus

The meaning of signs in sign language dictionaries and lexical databases is typically represented in terms of a spoken language, by including translation equivalents and sometimes also translated sentences illustrating typical use of signs. NGT Signbank lists translation equivalents in Dutch. At present, these translation equivalents are added based on the knowledge of annotators and researchers. These will often overlap with the meaning of actual uses of those signs in the corpus, but mismatches in both directions are observed: Signbank also lists translation equivalents that are not observed in the corpus, and not all possible translations of signs in the corpus are (yet) present in Signbank. These translations can be specified with each ID-gloss on a separate tier named Meaning in the corpus.

By harvesting the meaning annotations that are specified for ID-glosses, translation equivalents can be generated in Signbank in a corpus-based way.

4.2 Harvesting of mouthings in the corpus

The ubiquitous use of mouthings and their presumed role in the interpretation of NGT (Bank, 2015) calls for its systematic annotation in sign language corpora. The study of mouth actions in relation to signs continues to raise many questions. The mechanisms behind the variation found in the use of either mouthings or mouth gestures with signs, for instance, is not yet fully understood. Inclusion of corpus-based information on mouthings in the lexical database can help us to better understand the relation between manual signs and mouthings, and as for translation equivalents, frequency information on mouthings can aid in the determination of the semantics of signs.

One of the biggest challenges in the automated harvesting of mouthings, however, is temporal alignment. Mouthings do not necessarily align with the signs they accompany: they can spread over adjacent signs, or a sign can co-occur with multiple mouthings, and all the variations in between. Even when a stretch of connected signing co-occurs one-on-one with corresponding mouthings, annotation alignment is necessarily noisy, due to the complexity of the phonetic signal.

The solution we aim for is to list for each sign all mouthings that co-occur with that sign, including those that only partly overlap. In addition, two distinct values may be calculated and stored in relation to overlapping mouthings. First of all and most importantly, for each mouthing type, it should be calculated how often it occurs with a sign, just as for the translation equivalents discussed in the previous section. Second, the average amount of overlap of a mouthing type with a sign could be computed. The two numbers – frequency and overlap ratio – together provide a clear and concise measure of co-occurrence with sign types.

4.3 Use of corpus examples in the lexical database

A third possibility for enriching a corpus-based lexical database like Signbank would be to include information on the use of signs in their context. This can help in providing a richer view of the lexical semantics and pragmatics of signs, as well as form a solid basis for a learner dictionary in the long term (but see Hunston, 2009, on some of the complexities involved in presenting corpus data to learners).

5. Conclusion

In this paper we discussed several links between Corpus NGT annotations made in ELAN and the lexical database NGT Signbank. While the implementation of the links brings along some software development particular to the design of the two tools, the nature of the information is of a more general nature and has clear linguistic motivations. Information on lexical items stored in a lexical database is needed for a proper use of ID-glosses in the annotation of manual signs in sign language corpora. The frequency data, semantics and contextual information from corpora all form important additions to a lexical database. They can ultimately lead to corpus-based dictionaries (see also Hanke, 2006 for discussion).

The scenario we describe here is of course not the only one currently in use – but there are not too many alternatives. Hanke (2002) and Konrad & Langer (2008) describe the integrated iLex environment, where type and token data as well as metadata are integrated in a single database. This solution has also been adopted in Poland and Denmark, among other countries. Together with the scenario described in this paper, these two seem to be the only solutions world-wide that have a substantial number of users, both in terms of the sign languages covered and the number of research groups working with them.

6. References

Bank, R. (2015). *The ubiquity of mouthings in NGT. A corpus study*. Utrecht: LOT.

Crasborn, O., Bank, R., & Cormier, K. (2015). Digging into Signs: Towards a gloss annotation standard for sign language corpora. Nijmegen: Radboud University & London: University College London. http://www.ru.nl/publish/pages/723853/dis_annotation_guidelines_4may2015.pdf

Crasborn, O., Bank, R., Zwitserlood, I., Van der Kooij, E., Ormel, E., Ros, J., … Vonk, M. (2014). NGT Signbank. Nijmegen: Radboud University, Centre for Language Studies.

Crasborn, O., & Sloetjes, H. (2014). Improving the exploitation of linguistic annotations in ELAN. In *Proceedings of LREC 2014.*

Crasborn, O., & Zwitserlood, I. (2008). Annotation of the video data in the Corpus NGT. Nijmegen: Radboud University, Centre for Language Studies. http://www.ru.nl/publish/pages/527859/corpusngt_annotationconventions.pdf

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

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL Signbank : The lemma dilemma revisited. *International Journal of Lexicography*, *28*(2), 169-–206.

Hanke, T. (2002). *iLex - A tool for sign language lexicography and corpus analysis.* Paper presented at the LREC 2002 conference, Las Palmas de Gran Canaria, Spain.

Hunston, S. (2009). The usefulness of corpus-based descriptions of English for learners. The case of relative frequency. In Karin Aijmer (Ed.), *Corpora and language teaching* (Vol. 33, pp. 141-154). Amsterdam/Philadelphia: John Benjamin Publishing Company.

Johnston, T. (2001). The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics*, *4*(1), 145–169.

Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. In *Construction and exploitation of sign language corpora. Proceedings of the 3rd Workshop on the Representation and Processing of Sign Languages.*

Konrad, R., & Langer, G. (2009). Synergies between transcription and lexical database building: The case of German Sign Language (DGS). In M. Mahlberg, V. González-Díaz, & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference, CL2009*. Liverpool, UK.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006.*