Online Concordancer for the Slovene Sign Language Corpus SIGNOR

Špela Vintar, Boštjan Jerko

University of Ljubljana

Faculty of Arts, Aškerčeva 2, Ljubljana

E-mail: spela.vintar@ff.uni-lj.si, bostjan.jerko@guest.arnes.si

Abstract

We present the first version of an online concordancing tool for the Slovene Sign Language SIGNOR corpus. The corpus search tool allows querying the SIGNOR annotated database by glosses and displays the hits in a keyword-in-context (KWIC) format, accompanied by frequency information, HamNoSys transcription and metadata. The main purpose of the tool is linguistic research, more specifically sign language lexicography, but also providing general public access to the corpus.

Keywords: SIGNOR, sign language corpus, online concordancing, corpus search tool

1. Introduction

Slovene Sign Language (SZJ) is the primary language of the Deaf community in Slovenia comprising between 1000 and 1500 users. Within a 3-year research project, a corpus of SZJ was compiled by collecting video samples from 80 informants, which were then transcribed and annotated at several levels of analysis[1]. In this paper we present an online concordancing tool which can be used to query the corpus annotations, explore sign frequencies and view signs within the authentic conversational context.

Online corpus interfaces for sign language corpora are scarce. We are familiar with searchable sign databases such as the Lexical Database of Sign Language in Klagenfurt[2], the Auslan Signbank[3] (Johnston 2001) and the BSL Signbank[4] (Cormier et al. 2012), and the open access online corpus of movies representing Dutch Sign Language (NGT) (Crasborn and Slöetjes 2014). Sign databases are inventories of signs which do not provide contextual information and cannot replace sign language corpora, where authentic conversations have been recorded and annotated. The Dutch NGT corpus is based on the ELAN corpus annotation workbench [5] for multimodal corpora, and the multi-tier search functionality is provided by the TROVA search engine. The main problem with representing sign language in an online querying environment is the potential complexity of queries; sign language corpora typically contain multi-layered annotations where different types of data (glosses, timecodes, audiovisual data, metadata etc.) overlap and are difficult to present in a user-friendly manner. Furthermore, existing query tools rely on corpus annotation workbenches (ELAN or iLex) which are not easily portable into a web environment and usually require dedicated browsers.

Our aim was to create a simple web interface where the corpus could be searched from any browser, however our tool currently does not support complex or multi-layer queries.

2. The SIGNOR Corpus

The compilation of the corpus started in 2011. Preliminary considerations involved issues of regional balance, the informants' competence in SZJ, text types, communicative settings, and elicitation techniques, as well as technical issues regarding the recording sessions and video processing. Having reviewed several related projects, our methodology of video session organization relied on Nishio et al. (2010), and the segmentation and annotation strategies were also mostly adapted from the German DGS project (Hanke et al. 2012; Konrad et al. 2012).

All of the recordings were converted into a common data format and stored on the project data server. For corpus annotation we used the iLex tool (Hanke and Storz 2008), which provides a flexible multiuser annotation environment and stores all signs, lexemes, and tokens in a database, thus facilitating consistency between annotators.

Annotation includes the following layers (Vintar et al. 2012, Vintar 2015):

- Tokenization. The video stream of signed dialogue is segmented into individual signs delimited by time codes.
- Glossing. The process of assigning each sign a lexical identifier is also referred to as lemmatization; in other words, each token is assigned a type.
- Mouthing. The voiceless or voiced articulations accompanying signs may constitute, reinforce, or alter their meaning.
- Meaning. Each sign is assigned its meaning in the given textual context.
- Compound meaning. Many signs are compositional or phrasal, and the meaning of such multisign units is annotated as a separate tier.

[1] http://www.lojze.si/signor/index.html
[2] http://ledasila.uni-klu.ac.at/TPM/
[3] http://www.auslan.org.au/dictionary/
[4] http://bslsignbank.ucl.ac.uk/dictionary/
[5] http://tla.mpi.nl/tools/tla-tools/elan/

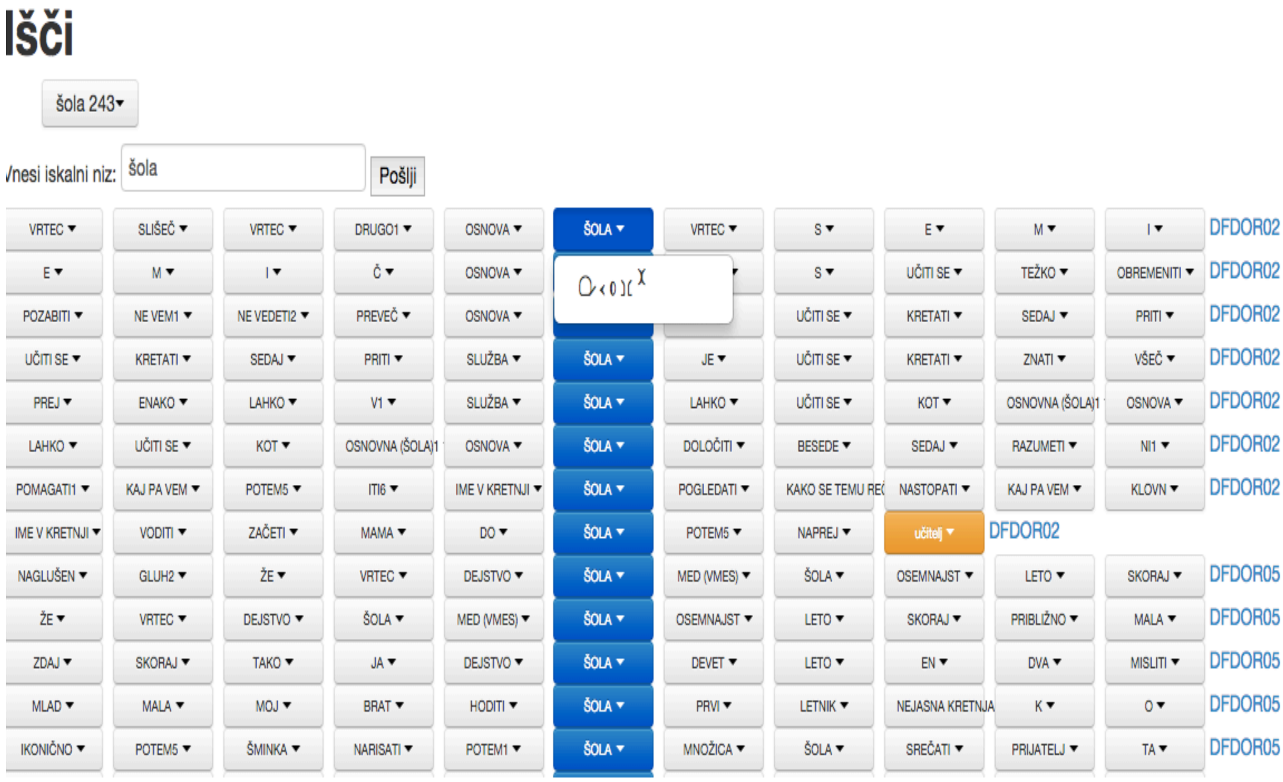

Figure 1: The SIGNOR search interface

- HamNoSys transcription (Schmaling and Hanke 2001). The graphical notation of signs helps distinguish sign variants and represents an important step for further processing or sign generation with animated agents.
- Segmentation into utterances. This step was performed on a section of the corpus comprising 3,000 utterances. Each utterance boundary is marked with a specific gloss indicating its form.

The overall length of recordings amounts to approximately 40 hours. The final size of the annotated corpus is 30,335 sign tokens and 2,976 sign types. Of the latter, 1,043 signs occur only once in our corpus. A lexical analysis of the corpus revealed that the frequencies of lexical categories roughly correspond to other sign language corpora (Vintar 2015). SZJ is rich in variants – up to 9 different variants have been found for the same sign, and it seems that variation occurs between different age groups, places of education and geographical regions.

3. The SIGNOR Concordancer

The aim of the search interface was to enable researchers, interpreters and SZJ users to explore signs in context and to compare the frequencies of various lexical items, including potential region- or age-related variants. The concordance line is composed of individual glosses, whereby compound signs are glossed with their complex meaning and marked in a different colour. For each concordance line the interface also displays the anonymized metadata upon click: Gender, Region, Level of Deafness, Education and Primary Hand of the informant.

As a main storage of all the data iLex uses PostgreSQL. For ease of access to the data we decided to access the database directly so we can automate the process of exporting annotations as needed. The concordancer uses MongoDB database for easier usage for online purposes. The data exported from PostgreSQL as CSV were imported to the concordancer database. The CSV files consist of an index of signs with associated data (start and end time codes, gloss, Hamnosys and compound meaning, if applicable). After that, a script is used to import all the data to MongoDB and another script to compute the frequencies of different signs.

The interface is simple and intuitive, providing a single search window to enter the query. The resulting concordance displays the search gloss in context in a keyword-in-context (KWIC) format, with a default window of +/-5 adjacent glosses. Compound meanings are written in lowercase and coloured orange so as to indicate that the sign is compositional. The frequency of the search gloss is displayed on top of the concordance window (Figure 2). A click on any gloss reveals the HamNoSys notation. If the user clicks the Hamnosys notation, an avatar is shown in a separate window signing the selected sign. We are currently using the avatar engine integrated into iLex (Figure 3).

Probably the most useful feature for the purposes of sign language lexicography, teaching or sociolinguistic research is the information on the frequency of sign variants. Thus, a query for AMERIKA ("America") will result in a KWIC display of all variants of the sign for AMERIKA, but the concordance can be filtered by

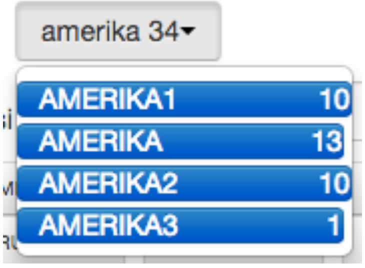

Figure 2: Sign variants

sign variants of which frequencies are displayed in a drop-down menu (Figure 2). Using filtering and the metadata links displayed for each concordance line, the user may draw conclusions on the distributional properties of each sign variant.

Figure 3: Sign animation generated from HamNoSys

4. Conclusion

This paper describes a simple online concordance tool for the SIGNOR corpus of Slovene Sign Language (SZJ). While advanced annotation tools such as ELAN or iLex allow for detailed and sophisticated queries of multimodal corpora, they are restricted to their own software environment and often too complex for the general public. Our purpose was to create an interface accessible to anyone, including Deaf people, sign language interpreters, teachers and students. It is also a good way of spreading the awareness about sign language among linguists and language policy makers. Our tool should be seen as work in progress as it has been developed within a very small nationally funded project, and the funding of future activities has not been secured yet.

Still, we plan to implement other features to better respond to the needs of potential users. One important future plan is to include the authentic video recordings into the online corpus, but currently we are still resolving legal issues related to data protection and have not obtained full permissions for the public release of all videos. Another improvement we plan is to include HamNoSys notations as a possible query type, so that users might have the possibility to search by signs or sign elements. Several technical improvements are also underway, including caching frequent searches for faster retrieval and optimizing for mobile access.

Acknowledgement

This work was partly funded by the Slovene Research Agency (ARRS) grant J6-4081, 2011-2014.

Bibliography

Cormier, Kearsy, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam, and Bencie Woll. 2012. From corpus to lexical database to online dictionary: Issues in annotation of the BSL Corpus and the development of BSL SignBank. *5th Workshop on the Representation of Sign Languages: Interactions between Corpus and Lexicon [workshop part of 8th International Conference on Language Resources and Evaluation*, Turkey, Istanbul LREC 2012. Paris: ELRA. pp. 7–12.

Crasborn, Onno, and Han Slöetjes. 2014. Improving the Exploitation of Linguistic Annotations in ELAN. In N. Calzolari, K. Choukri & et al. (Eds.), *Proceedings of LREC 2014, Ninth International Conference on Language Resources and Evaluation*. Paris: ELRA.

Hanke, T., and J. Storz. 2008. iLex: A Database Tool For Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In *Proceedings of the Language Resources and Evaluation Conference 2008, May 28–May 30.* Paris: ELRA.

Hanke, T., S. Matthes, A. Regen, and S. Worseck. 2012. Where Does a Sign Start and End? Segmentation of Continuous Signing. In *Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC), Istanbul,* ed. O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch, 69–74. Paris: ELRA.

Johnston, Trevor. 2001. The lexical database of Auslan (Australian Sign Language). *Sign Language & Linguistics* 4.1-2 (2001): 145-169.

Konrad, R., T. Hanke, S. König, G. Langer, S. Matthes, R. Nishio, and A. Regen. 2012. From Form to Function: A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC), Istanbul,* ed. O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch, 87–94. Paris: ELRA.

Nishio, R, S.-E. Hong, S. König, R. Konrad, G. Langer, T. Hanke, and C. Rathmann. 2010. Elicitation

Methods in the DGS (German Sign Language) Corpus Project. In *Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies,* 178–85. Paris: ELRA.

Schmaling, C., and T. Hanke. 2001. *HamNoSys 4.0*. http://www.sign-lang.uni-hamburg.de/Projekte/HamNoSys/HNS4.0/englisch/HNS4.pdf, accessed October 13, 2014.

Sloetjes, H., and Wittenburg, P. 2008. Annotation by category – ELAN and ISO DCR. In: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).* Paris: ELRA.

Vintar, Š., Jerko, B. and Kulovec, M. 2012. Compiling the Slovene Sign Language Corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon Language Resources and Evaluation Conference (LREC), Istanbul,* ed. O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Kristoffersen, and J. Mesch, 87–94. Paris: ELRA.

Vintar, Š. 2015. Lexical Properties of Slovene Sign Language: A Corpus-Based Study. *Sign Language Studies*, *15*(2), pp.182-201.