Towards tagging of multi-sign lexemes and other multi-unit structures

Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer

University of Hamburg
Institute of German Sign Language and Communication of the Deaf
Binderstr. 34, 20146 Hamburg, Germany
E-mail: {thomas.hanke, susanne.koenig, reiner.konrad, gabriele.langer}@sign-lang.uni-hamburg.de

Abstract

With the building of larger sign language corpora tagging, handling and analysing large amounts of data reach a new level of complexity. Efficiency and interpersonal consistency in tagging are relevant issues as well as procedures and structures to identify and tag relevant linguistic units and structures beyond and above the manual sign level. We present and discuss problems and possible solution approaches (focussing on the working environment of iLex) of how to deal with multi-unit structures and more specifically multi-sign lexemes in annotation and lexicon building.

Keywords: sign language tagging, multi-word expressions, annotation tools

1. Multi-sign expressions as lexicon entities

For many sign languages, compounds and idiomatic expressions are attested to occur. Becker (2003) differentiates between proper compounds and loan compounds in DGS (German sign language), with the first group being rather rare. Johnston & Ferrara (to appear) report that multi-sign idiomatic expressions are rare as well.[1] However, some of these units may have not been discovered yet since empirical studies on these topics require large amounts of data, which are only now becoming available with the development and accessibility of large sign language corpora. In addition, often there are no clear-cut distinctions between the three groups mentioned.

Whatever the exact definitions for these phenomena are, they have something in common: When these patterns appear, they are different from just the signs they consist of. There may be restrictions on the use of these patterns not be expected from how their components can be used, and the meaning might be different from the composition of meanings of their building blocks, or they might disambiguate POS attributions to their parts. This means they have to be considered part of the language's lexicon. Once multi-unit structures are stored in the lexical database, they can also be attributed with all kinds of lexicographic annotation such as regional use or syntactic restrictions. This would of course also be the place to further characterise the construction, e.g. what kinds of variation does it allow.

2. Tagging multi-sign expressions

In today's coding conventions (e.g. Johnston 2011, but also including our own), these multi-unit structures are not really dealt with in a way that allows to mark, access, list and describe these units as entities of their own right. Usually, only their building blocks are made visible in the annotation. Occurrences of these structures need to be retrieved by executing searches. This is unsatisfactory not only in the context of complex patterns difficult to search for, but also from the point of view of lemma revision. Searches will retrieve all occurrences of the patterns, and the information whether the pattern is actually used in the special (e.g. idiomatic) sense or in the literal that is sign-by-sign sense is not stored anywhere. We therefore look for a possibility to clearly identify tokens of multi-sign structures while maintaining the tagging of the constituents.

A first approach would be to have a separate tier tagging those time stretches where multi-sign structures occur with labels such as "idiom" or "compound". To find the multi-sign instances, one would then set up the search not only containing the sign pattern, but also the extra label. Nesting structures would require multiple extra tiers which is uncomfortable but still manageable. This approach, however, does not generalise to other forms of multi-unit structures such as multi-channel signs (e.g. a sign with an obligatory lexicalised mouth gesture) or discontinuous structures such as sandwich verbs or resumed holds: A simple tag in a separate tier would either include all co-occurring events in a tier or none.

An alternative actually in use for spoken languages multi-word expressions is to store them as the pre-terminal level in a treebank. However, today there is no annotation system for sign languages featuring treebanks.

What we currently envision is to add this lowest level of syntax trees to the tiers & tags model of iLex[2]. As we would not expect more than two levels above the basic token tags, these extra levels would be projected onto the iLex annotation grid display by framing those tags

[1] Their claim is for idiomatic expressions in Auslan, but the same seems to be the case for DGS.

[2] iLex is the transcription environment we use which features a lexical database closely integrated with the annotation scores, cf. Hanke (2002) and Hanke & Storz (2008).

constituting a multi-unit structure. Token structures compatible with the multi-unit structure as stored in the lexicon can be claimed an instance of that structure by dragging the lexical item onto one of the existing tokens.

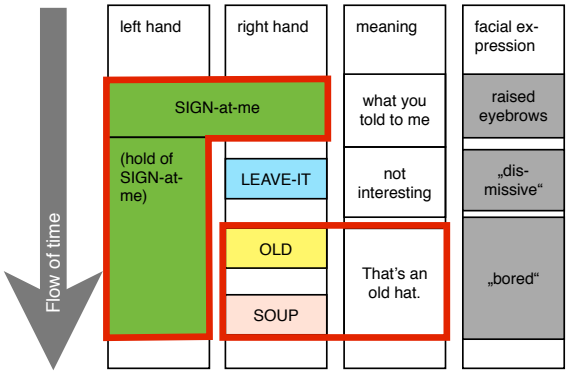

Figure 1: Idiomatic expression and hold structure

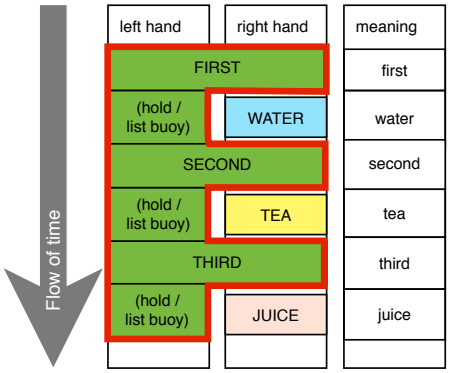

Figure 2: Multi-item list structure

3. Multi-sign lexemes in the lexical database

Naturally, this approach requires the lexicon structure in iLex to be extended in order to cover multi-sign multi-channel structures within the lexical database. Currently, a simplex sign is described as either one- or two-handed, with an optional code for mouthing or mouth gesture that may be copied to the mouth tier but is not considered part of the token. Complex signs are either simultaneous or sequential compounds or blends of two simplex signs. More complex structures cannot be appropriately handled in the implemented lexicon model. The idea is to allow any kind of element (simple signs, nonmanuals etc.) to be arranged in a structure expressing time relations such as "precedes" or "precedes immediately". To the user, these structures would appear as miniature transcripts without concrete timestamps.
An extension allowing multi-channel signs would also cover obligatory facial actions for lexemes in a much more transparent way than the current solution.

4. Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Programme, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Programme is coordinated by the Union of the German Academies of Sciences and Humanities.

5. References

Becker, C. (2003). *Verfahren der Lexikonerweiterung in der Deutschen Gebärdensprache*. Hamburg: Signum.

Hanke, T. (2002): iLex. A tool for Sign Language Lexicography and Corpus Analysis. In González Rodriguez, M. & Paz Suarez Araujo, C. (Eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas de Gran Canaria, Spain*; Vol. III. Paris: ELRA, pp. 923-926.

Hanke, T., Storz, J. (2008). iLex - A Database Tool for Integrating Sign Language Corpus Linguistics and Sign Language Lexicography. In Crasborn, O., Efthimiou, E., Hanke, T., Thoutenhoofd, E., Zwitserlood, I. (Eds.), *LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora.* Paris: ELRA, pp. 64-67. Available online at http://www.lrec-conf.org/proceedings/lrec2008/workshops/W25_Proceedings.pdf

Johnston, T. (2011). *Auslan Corpus Annotation Guidelines. 30. November 2011*. Available online at http://www.auslan.org.au/video/upload/attachments/AuslanCorpusAnnotationGuidelines30November2011.pdf

Johnston, T., Ferrara, L. (in press). Lexicalization in Signed Languages. When is an Idiom not an Idiom ? Online Proceedings of UK-CLA Meetings [Proceedings of the 3rd UK Cognitive Linguistics Conference, University of Hertfordshire: 6-8 July 2010.] Available online at http://mq.academia.edu/TrevorJohnston/Papers/902380/Lexicalization_in_signed_languages_when_is_an_idiom_not_an_idiom