Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing

Gabriele Langer1, Thomas Troelsgård2, Jette Kristoffersen3, Reiner Konrad4, Thomas Hanke5, Susanne König6
1University of Hamburg, Institute for German Sign Language and Communication of the Deaf, Germany
2 Center for Sign Language, University College Capital, Copenhagen, Denmark

Introduction and Background

- Cooperation of two projects: DTS Dictionary (Copenhagen, Denmark)
- DGS Corpus (Hamburg, Germany)
- Common goals: corpus building & corpus-based lexicography
- Tool: Lex
- Different directions & backgrounds: DGS Corpus: having corpus, approaching dictionary-making
- DTS Dictionary: having dictionary, approaching corpus-building
- Need: lexical database that facilitates corpus annotation as well as dictionary editing

Structural Requirements

Annotation and lexicographic description of signs serve different purposes and may model signs differently.

Lexicographic Needs

- Consistently annotated corpus data as basis for a sound lexicographic description of signs. Data serves as empirical evidence.
- Main focus is on meaning, including secondary and roughly distinguished.
- Word forms, phonological variants and type hierarchies with several levels for modeling differences in iconicity, form, and use of a sign and tagging the tokens accordingly. Not the same as lemmatisation (collocations, grammatical functions and usage). Overview of forms (including types for dictionary editing do not serve interlinked secondary and roughly distinguished.

Dictionary Entries

- Type entries for token-type matching (lemmatisation): search options: easy to find and identify adequate type/cohorts
- Direct access to citation form and/or representative token
- Direct access to all tokens of the respective type (for comparison and lemma revision)
- New preliminary types can be added as needed
- Provide a place for tokens with ambiguous, unclear or unusual (productive) contextual meanings (grouped with other tokens of the same form)
- For data analysis: queries, views, statistics, maps

Corpus Needs

- Type entries for token-type matching (lemmatisation)
- Main focus is on meaning, being secondary and roughly distinguished.
- Word forms, phonological variants and type hierarchies with several levels for modeling differences in iconicity, form, and use of a sign and tagging the tokens accordingly. Not the same as lemmatisation (collocations, grammatical functions and usage). Overview of forms (including types for dictionary editing do not serve interlinked secondary and roughly distinguished.

Dictionary Editing

- For the description of signs in dictionary entries other groupings may be more suitable.
- Establishing lemma signs for dictionary entries focus more on semantic aspects and usage.
- One annotation supertype may be split into several dictionary entries: Example: STAMP1-$SAM includes phonological variants A, B and modified forms C, D, E. Example: WORK2-$SAM covers word forms E, F, G, H and L.

Annotation and lexicographic description of signs serve different purposes and may model signs differently. Two different structures are needed: The level 2 subtype supertypes represent established or conventional uses of a sign with regard to variant forms and meaning (pre-sorting for roughly meanings (Hamburg) or identified senses (Copenhagen)).

References
