How Much Top-Down and Bottom-Up do We Need to Build a Lemmatized Corpus?

Susanne König, Reiner Konrad, Gabriele Langer, Rie Nishio
University of Hamburg, Institute of German Sign Language and Communication of the Deaf

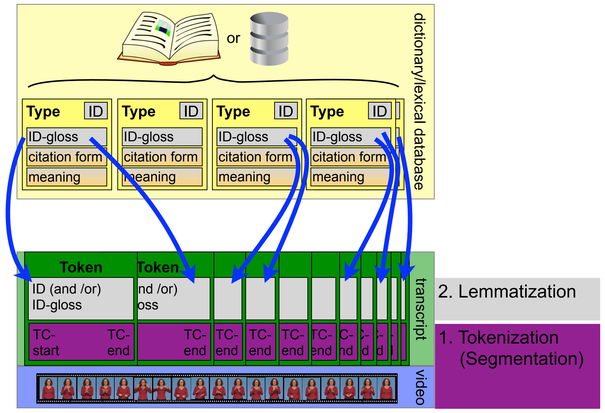
Lemmatization (Token-Type Matching) as a Strict Top-Down Process

Gloss annotation as procedure for tokenizing and lemmatizing sign language data
(corresponding to standard lemmatizing prodecure for written corpora with semi-automated processing)

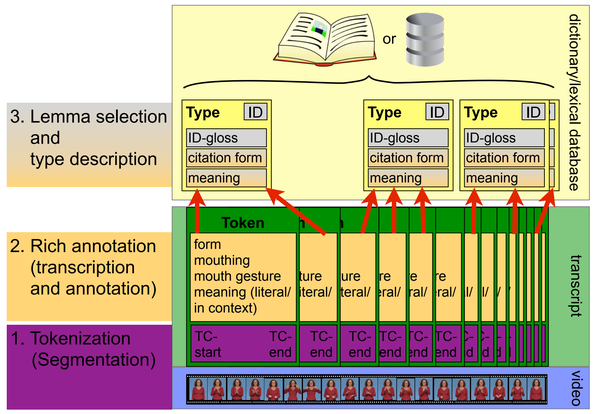
Prerequisite:
- comprehensive dictionary or lexical database
- efficient retrieval functions

Pros:
- highly consistent
- no transcription needed

Cons:
- no token information (form, meaning, …) available
- tokens of semi- and non-lexicalized forms (e.g. productive signs) cannot be matched
- existing lemma selection leads to a gap between corpus evidence and pre-defined categories



Lexicon Building as a Strict Bottom-Up Process



Ideal approach for spoken/signed languages without written form

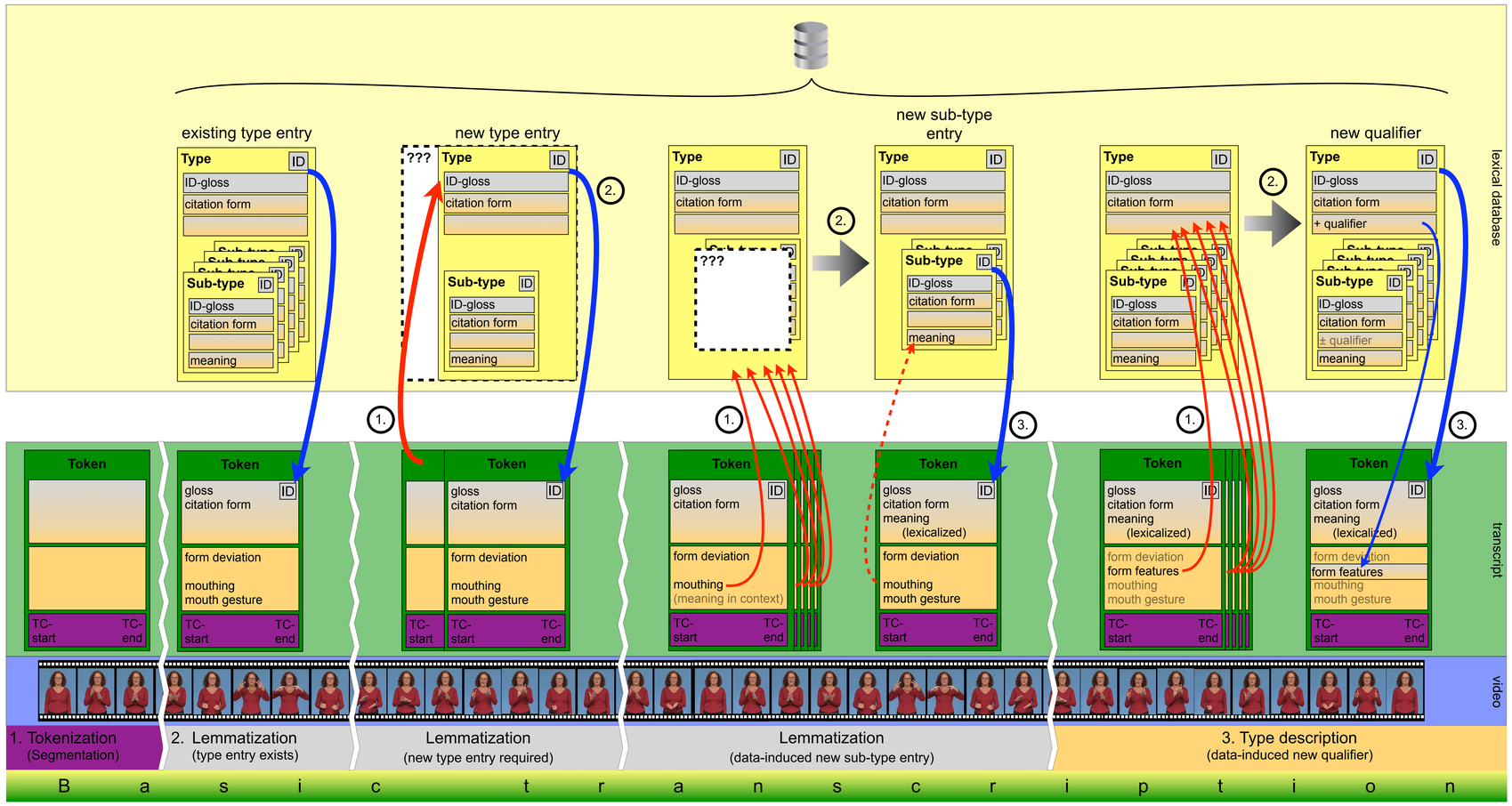
Prerequisites:
- richly annotated and time-aligned reference corpus
- precise annotation guidelines (transcription manual)
- additional processes and search routines to retrieve and group all tokens of one type after first round of annotation (lemmatization)
- clear criteria for lemma selection

Pro:
- data-driven
  - criteria for lemma selection are induced from the data
  - data not only confirms pre-existing hypotheses, but provides new information (heuristic value)
  - appropriate representation of language in use
  - semi- and non-lexicalized forms are captured

Cons:
- highly time-consuming
- prone to be inconsistent

Lemmatization and Lexicon Building in the DGS Corpus Project Using iLex

Time-Saving Top-Down Approach that is Continuously Counterbalanced by Token Information (Bottom-Up)



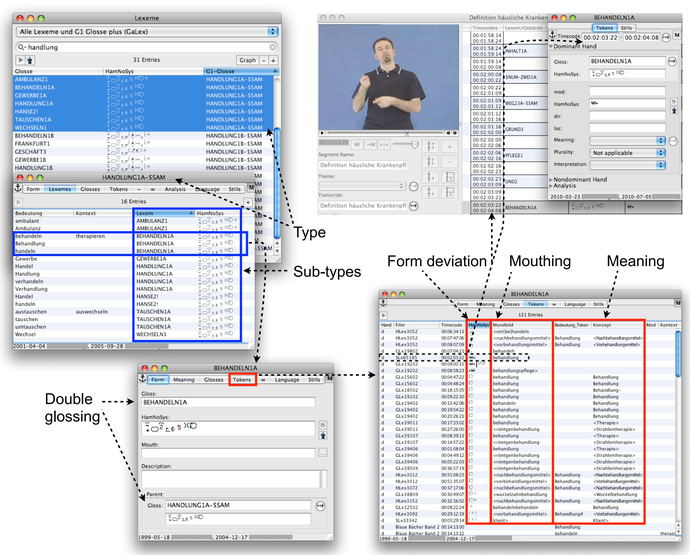
Research Objectives of the DGS Corpus Project:

- building the first lemmatized and annotated reference corpus of DGS
- compiling an electronic corpus-based dictionary of DGS

To Build a Lemmatized Corpus We Need:

- native signers as co-workers as well as feedback from the deaf community (e.g. voting via web-interface)
- a lexical database that
  - supports consistent token-type matching in a multi-user mode
  - allows for distinguishing between conventional and productive sign-mouthing combinations (until now realized by a hierarchical model of types and sub-types using **double glossing**)
  - allows for classifying form-function units (qualifiers; aka modifications)
  - supports **lemma revision** by comparing relevant token information of all tokens of one type (retrieval, listing, and sorting functions)
- explicit transcription guidelines for basic and detailled transcription
- minimal information on the form of a token (deviation from citation form)
- minimal information on the meaning of a token (mouthing, meaning of loan translations or context meaning)

Lemma Revision: Data Access in iLex



Work in Progress:

There is no comprehensive dictionary of DGS available for a strict top-down approach. However, a large pool of sign entries from previous LSP dictionary projects and published DGS compilations is available in iLex. These entries can be used for top-down procedures. The lexical database is continuously expanded via bottom-up procedures.

Poster presented at the Theoretical Issues in Sign Language Research (TISLR) 10 Conference, Sept 30 - Oct 2, 2010 at Purdue University, Indiana, USA. The research leading to these results has received funding from the German Academies of Science Programme.