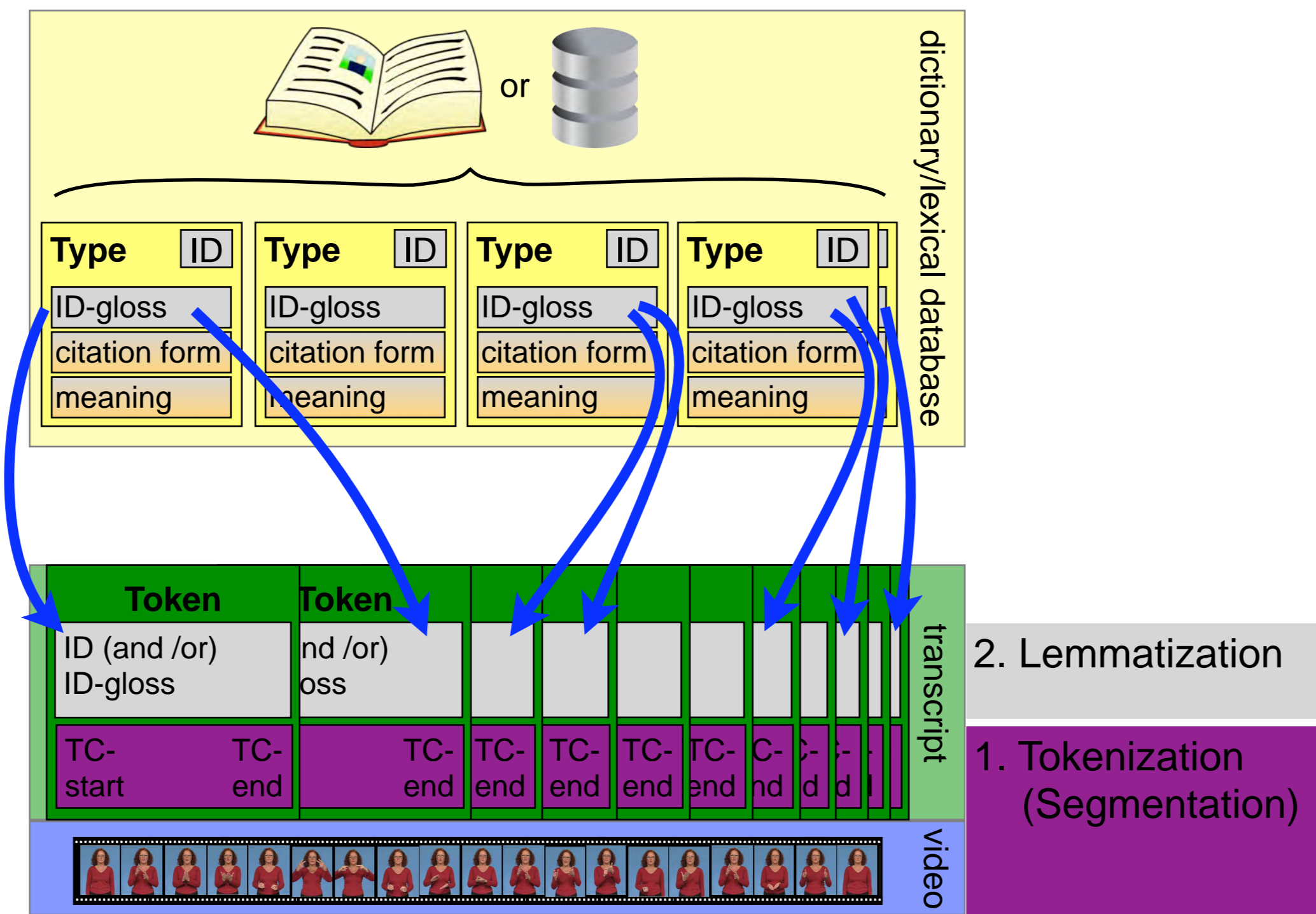# How Much Top-Down and Bottom-Up do We Need to Build a Lemmatized Corpus?

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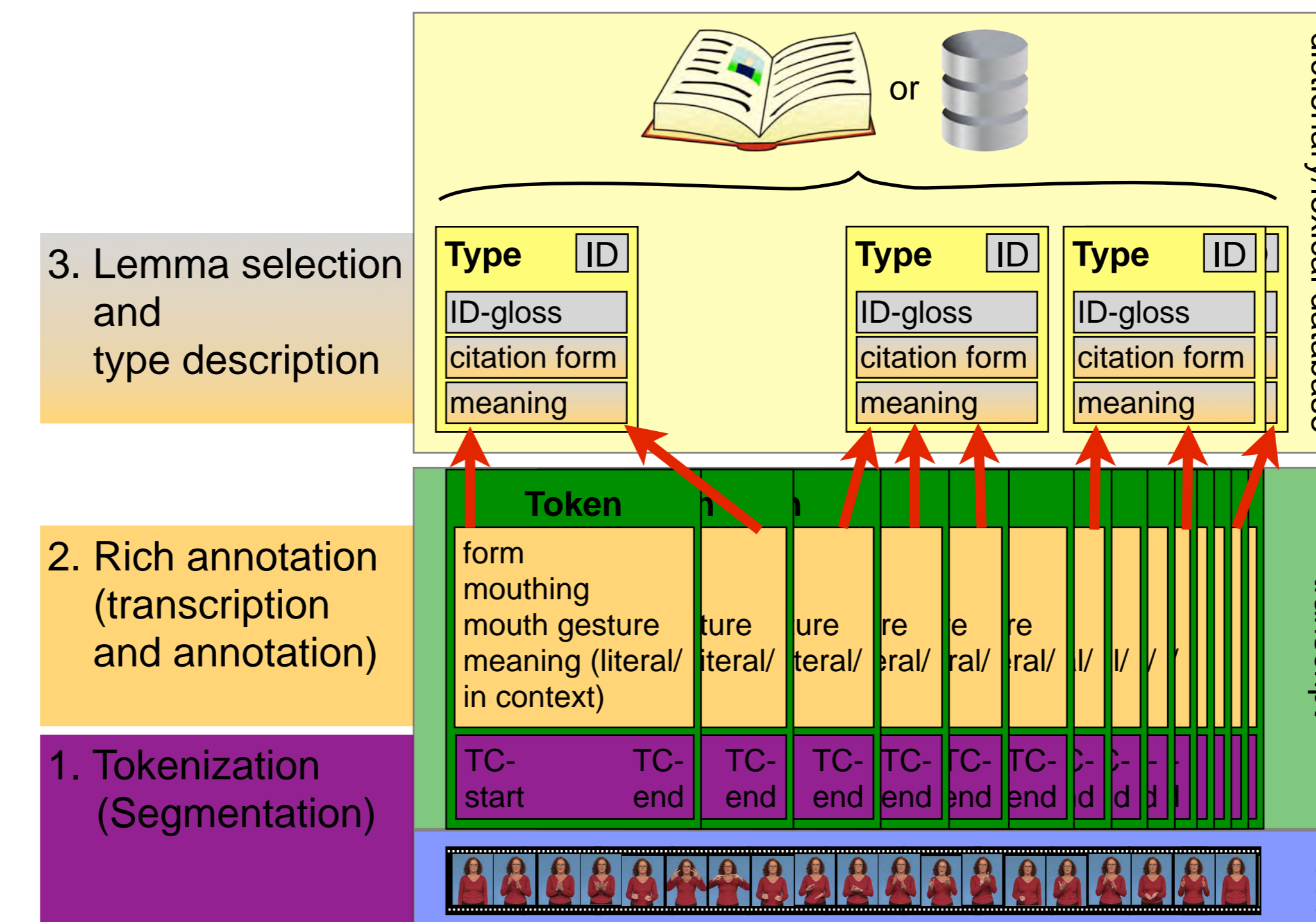- comprehensive dictionary or lexical database
- efficient retrieval functions

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

(diagram: dictionary/lexical database — Type / ID / ID-gloss / citation form / meaning; transcript — Token / ID (and /or) ID-gloss; 2. Lemmatization; 1. Tokenization (Segmentation); TC-start / TC-end; video)

## 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)
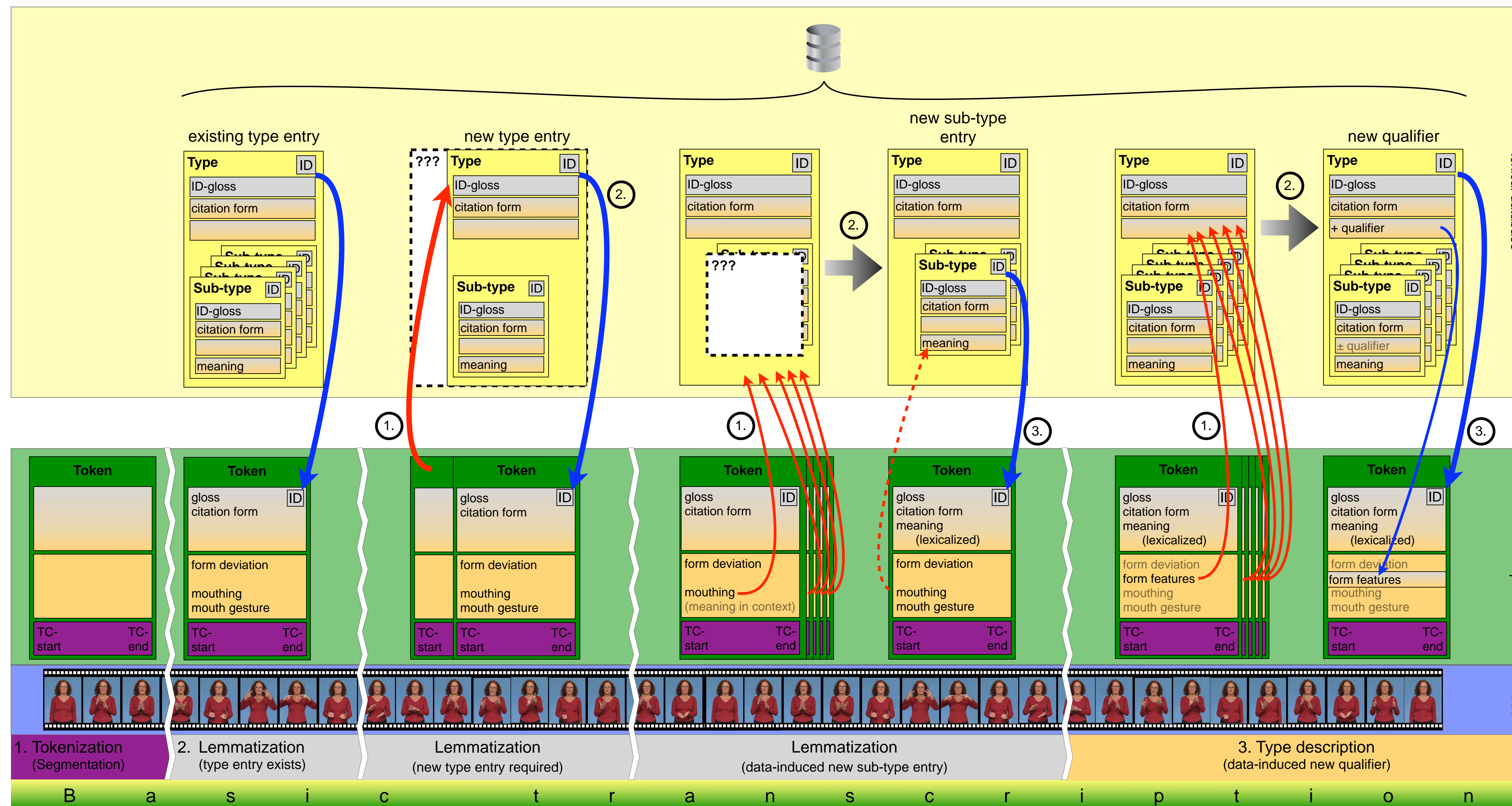- clear criteria for lemma selection

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

(diagram: 3. Lemma selection and type description; 2. Rich annotation (transcription and annotation) — form / mouthing / mouth gesture / meaning (literal/ in context); 1. Tokenization (Segmentation); dictionary/lexical database — Type / ID / ID-gloss / citation form / meaning; transcript; TC-start / TC-end; video)

## 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)

(large diagram with lexical database entries):
- existing type entry — Type / ID / ID-gloss / citation form / Sub-type / ID / ID-gloss / citation form / meaning
- new type entry — ??? / Type / ID / ID-gloss / citation form / Sub-type / ID / ID-gloss / citation form / meaning
- (Type / ID / ID-gloss / citation form / ??? )
- new sub-type entry — Type / ID / ID-gloss / citation form / Sub-type / ID / ID-gloss / citation form / meaning
- Type / ID / ID-gloss / citation form / Sub-type / ID / ID-gloss / citation form / meaning
- new qualifier — Type / ID / ID-gloss / citation form / + qualifier / Sub-type / ID / ID-gloss / citation form / ± qualifier / meaning

(transcript Tokens): gloss / citation form / ID / form deviation / mouthing / mouth gesture / TC-start / TC-end; meaning (lexicalized); meaning (meaning in context); form features

- 1. Tokenization (Segmentation)
- 2. Lemmatization (type entry exists)
- Lemmatization (new type entry required)
- Lemmatization (data-induced new sub-type entry)
- 3. Type description (data-induced new qualifier)
- B a s i c   t r a n s c r i p t i o n

### 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 detailed 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

(screenshots labeled: Type / Sub-types / Double glossing / Form deviation / Mouthing / Meaning)
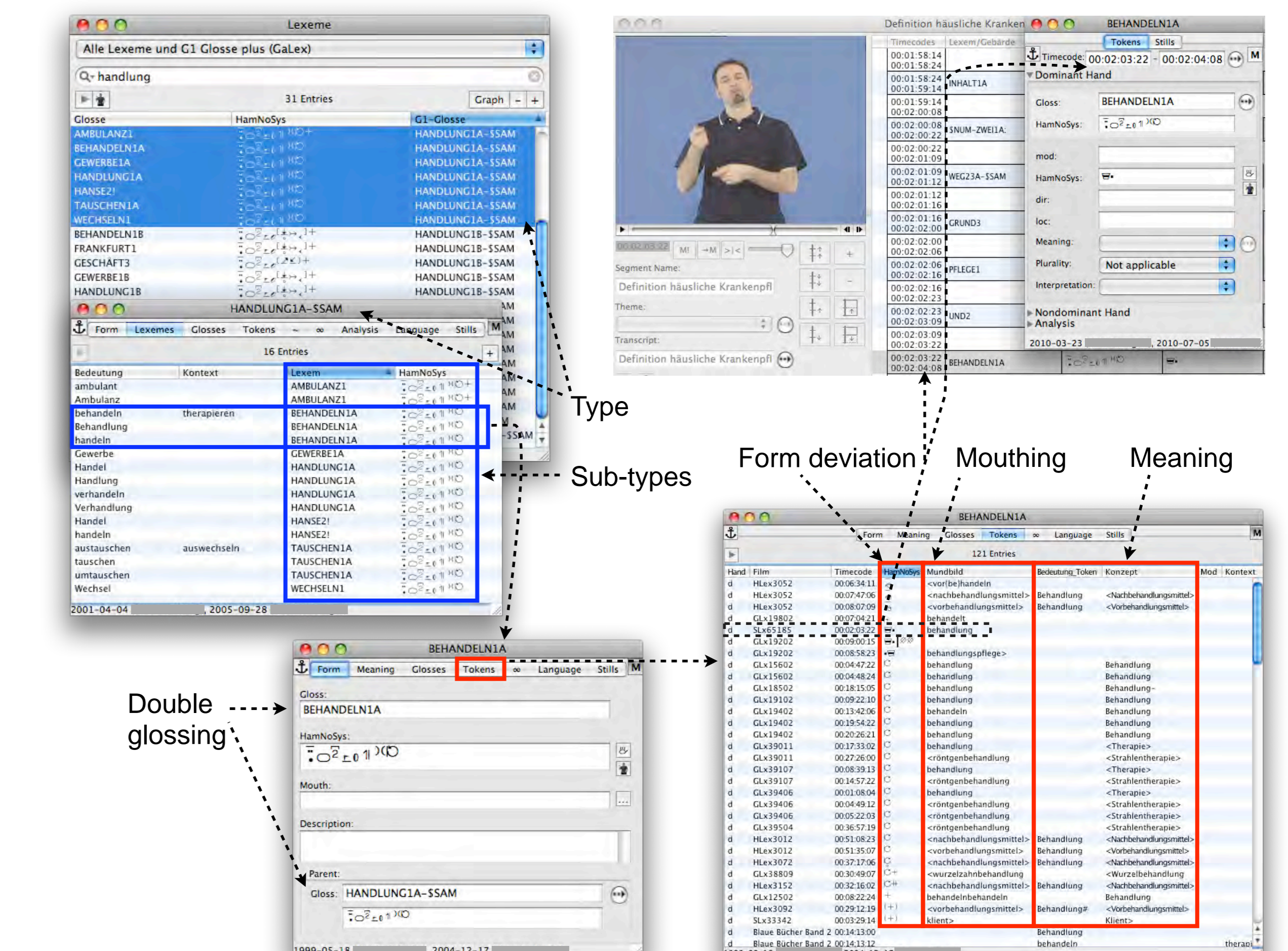
## Work in Progress:
There is no comprehensive dictionary of DGS available for a strict top-down approach. However, a large pool of sign entries from previous LSP dictionary projects and published DGS compilations is available in iLex. These entries can be used for top-down procedures. The lexical database is continuously expanded via bottom-up procedures.