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

Gabriele Langer[1], Thomas Troelsgård[2], Jette Kristoffersen[2], Reiner Konrad[1], Thomas Hanke[1], Susanne König[1]

[1] University of Hamburg, Institute for German Sign Language and Communication of the Deaf, Germany

[2] Center for Sign Language, University College Capital, Copenhagen, Denmark
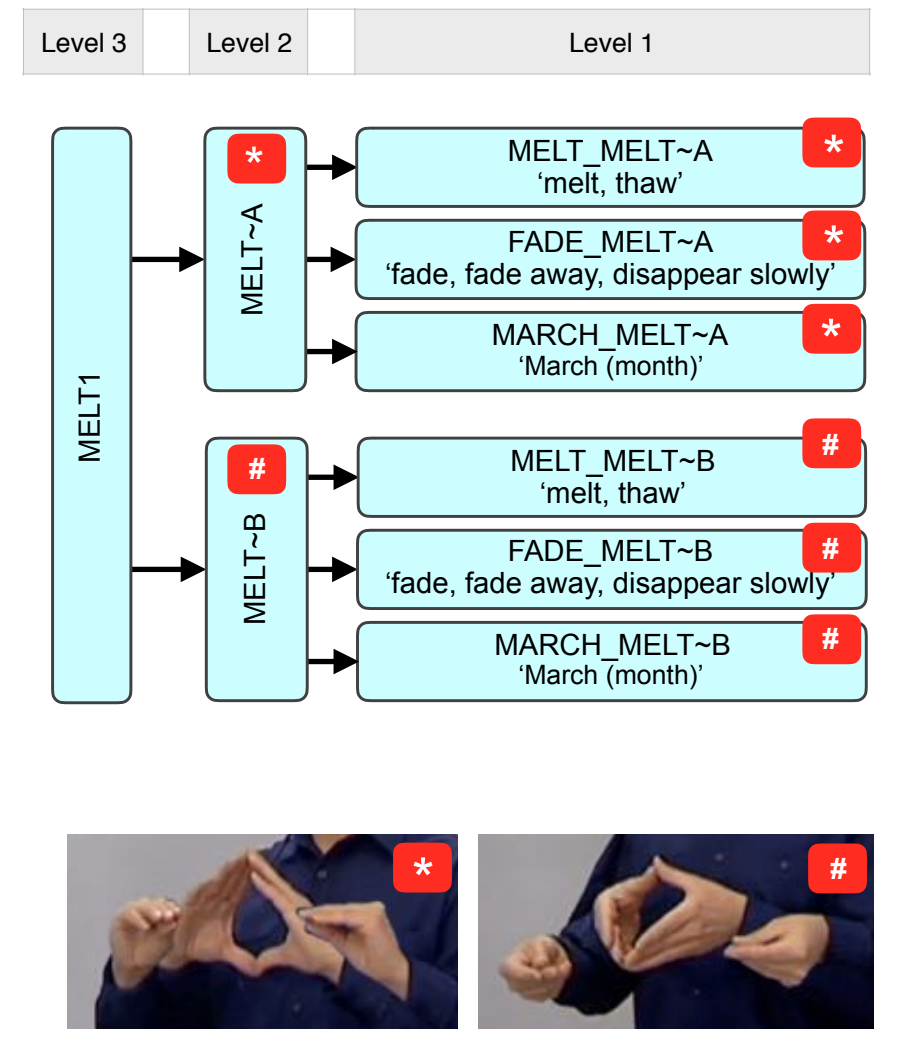
Introduction and Background

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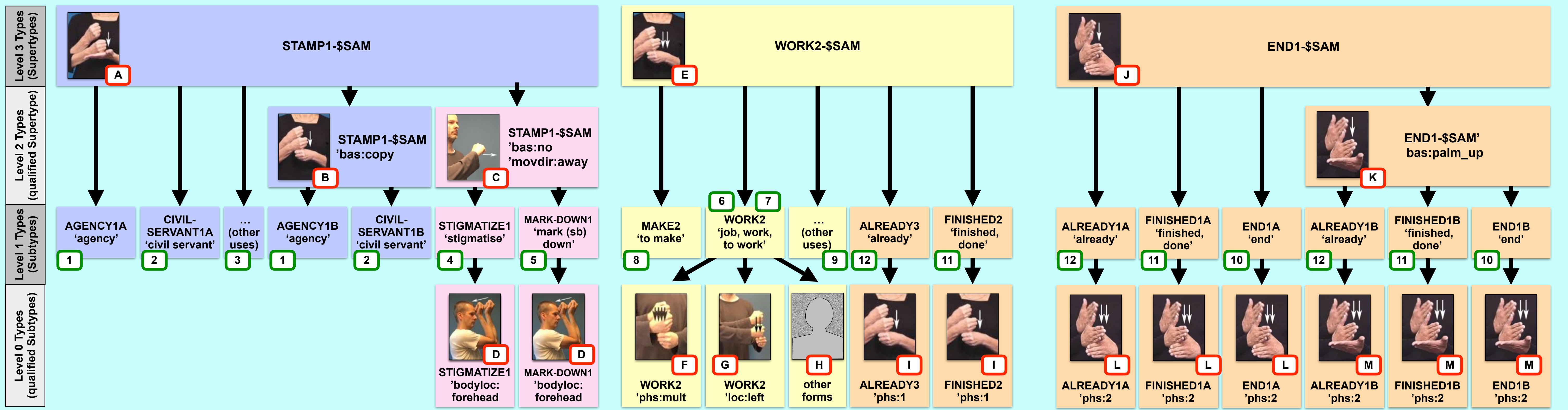- Need: lexical database that facilitates corpus annotation as well as dictionary editing
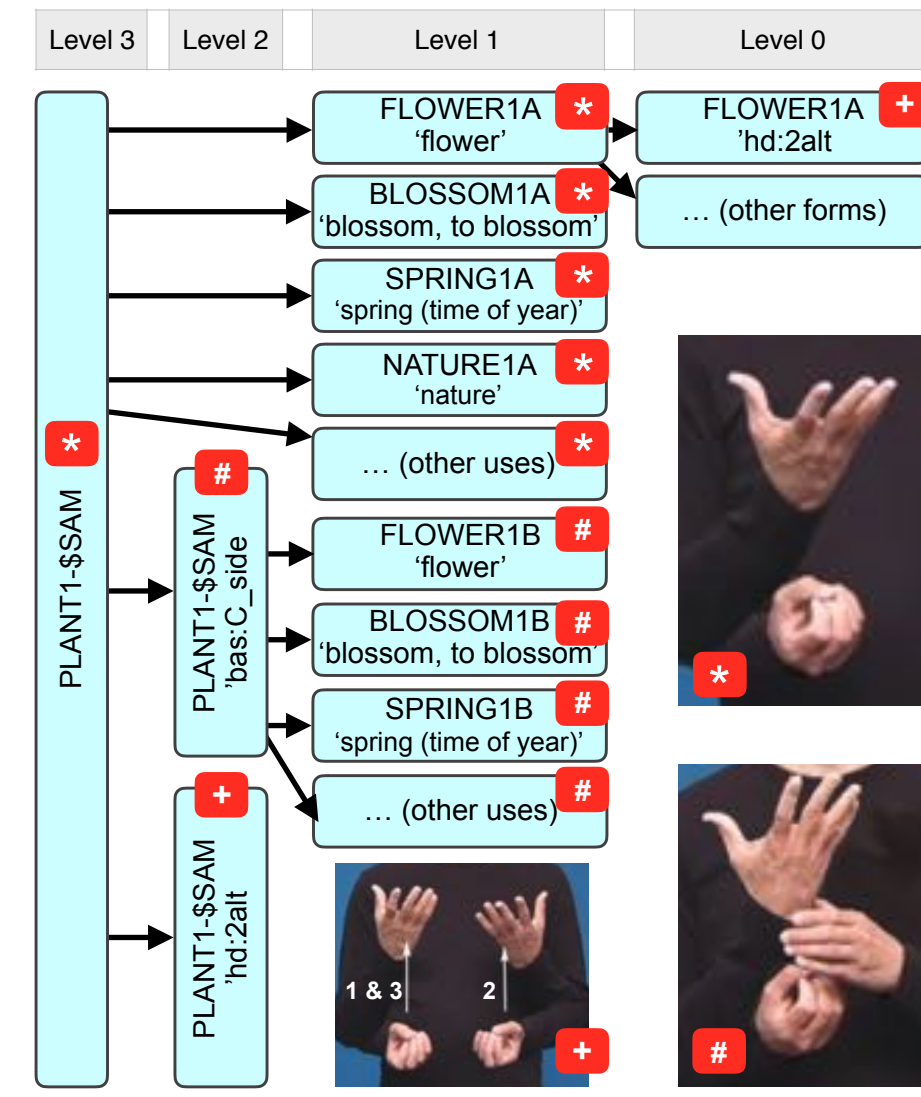
Example of type structure in iLex as used by DGS Corpus Project

iLex

- Database and annotation environment (Hanke 2002)
- Videos and time-aligned annotations
- Multi-user-approach: real-time access to new information by all users
- Inventory of types (type entries) for lemmatisation
- Type hierarchies with several levels for modeling relevant differences in iconicity, form, and use of a sign and tagging the tokens accordingly
- Different but similar use of type hierarchy structures by both projects (see examples left and right of this box)
  - Level 3: *supertypes* – represent signs as abstract linguistic units
  - Level 1: *subtypes* – 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))

Example of type structure in iLex as used by DTS Dictionary Project

Dictionary Entries

Explanation of Examples

Types for corpus annotation and types for dictionary editing do not match in a straightforward way.

Corpus Annotation

Annotation types focus mainly on form: one supertype includes phonological variants, modified forms and other word forms.
Example: STAMP1-$SAM includes phonological variants A, B and modified forms C, D.
Example: WORK2-$SAM covers word forms E, F, G, H and I.

Corpus Needs

- Type entries for token-type matching (lemmatisation).
- Main focus is on form, meaning being secondary and roughly distinguished.
- Word forms, phonological variants and presumably derived forms are grouped together (in iLex: belonging to same supertype).
- Meaning is secondary for identification of types (constituting subtypes in iLex).
- Annotation is a process in flux.
- Database/Annotation environment should facilitate efficient and consistent corpus annotation (lemmatisation):
  - search options: easy to find and identify adequate types/glosses
  - 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 same form)
  - for data analysis: queries, views, statistics, maps.

Lexicographic Needs

- Consistently annotated corpus data as basis for a sound lexicographic description of signs. Data serves as empirical evidence.
- Main focus of description is on meaning and usage. Overview of forms (including phonological variants and modified forms), meanings, usage (collocations, grammatical functions and patterns), and distribution (with regard to region, age, gender etc.).
- Establishment of lemma signs: consideration of forms, meanings, and usage. Further criteria defined by editorial rules (including pragmatic decisions). Not the same as lemmatisation (cf. Svensén 2009, 94).
- Lexicographic description is a summary of the data at one point in time (frozen status quo).
- Database/Lexicographic workbench should facilitate the analysis and editing process:
  - establishment of entries (lemma signs)
  - tools for analysis (e.g. maps, statistics)
  - concordance view of tokens in context
  - provide a place for a preliminary fine-grained description of occurring senses
  - link to or include empirical evidence and examples backing all pieces of information on the sign
  - provide all structural elements of final entry
  - workflow control, status marker, and the like.
- Aim: Producing dictionary entries that are human-readable as well as fairly homogeneous in appearance.

Structural Requirements

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

➔ Two different structures are needed:
  - sufficiently similar to be related to each other
  - to be interlinked with each other
  - relatively free linking from dictionary editing structures to all levels of annotation database and corpus data
  - dictionary links to evidence and examples in corpus.
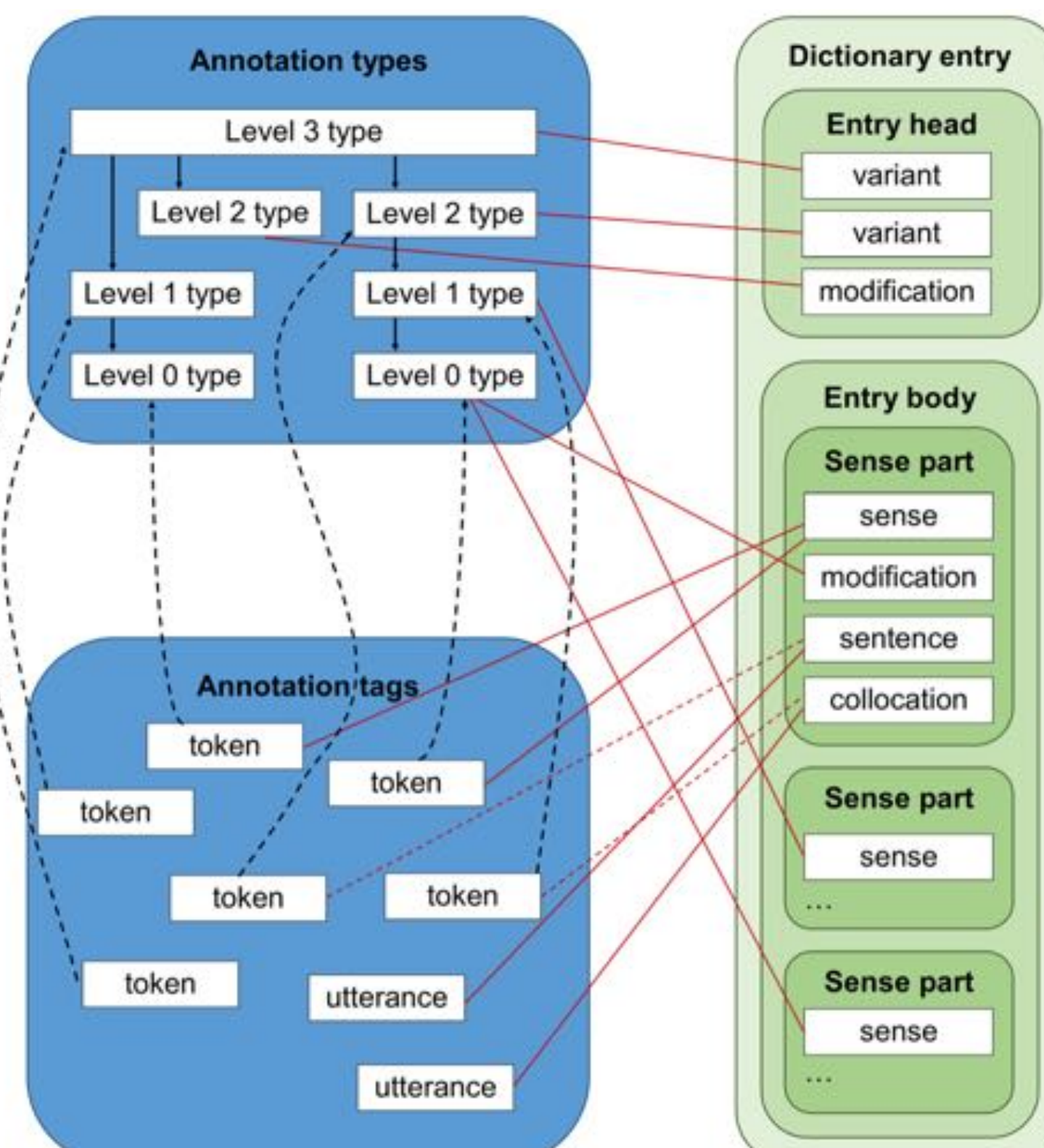
Dictionary Editing

- 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-$SAM is split into entries STAMP: sign forms A and B and senses like 'agency', 'civil servant' etc. (1, 2, 3), and MARK-DOWN: sign forms C, D and senses 'stigmatise' (4) and 'mark somebody down' (5).
- One dictionary entry may cover only part of the data of the annotation supertype:
  Example: Entry WORK2 (forms E, F, G, H and senses 6, 7, 8, 9) does not cover all data of WORK2-$SAM.
- One dictionary entry may combine material from different annotation supertypes:
  Example: From a form-based perspective the sign forms E and I, belong to supertype WORK2-$SAM (which also includes the forms F, G, H). However, concerning their meanings and usage they are best described together with forms J, K, L and M of the supertype END1-$SAM in their own dictionary entry (FINISHED), because they share the senses 'finished, done' (11) and 'already' (12), and because their manual form only varies in handshape (and palm orientation of the passive hand) and could be considered phonological variants of the forms J, K, L and M.

References
Center for Tegnsprog (2008-2016). *Ordbog over Dansk Tegnsprog*. Available at: www.tegnsprog.dk.
DGS Corpus Project. www.dgs-korpus.de.
Hanke, T. (2002): iLex. A tool for Sign Language Lexicography and Corpus Analysis. In M. González Rodriguez, & C. Paz Suarez Araujo (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. [Online resource; URL: www.lrec-conf.org/proceedings/lrec2002/pdf/330.pdf ; last access: March 22, 2016].
Svensén, B. (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.