

Workflow of Corpus-Based Lexicography

Deliverable COST-ENeL-WG3 meeting

July 2014, Bolzano/Bozen

Carole Tiberius & Simon Krek

1 Introduction

The focus of the Bolzano meeting was on the workflow of corpus-based lexicography. Up to now, research into the lexicographical workflow/process has mainly been carried out and described for printed dictionaries (Dubois 1990; Landau 1984; Zgusta 1971). Three phases are generally distinguished (e.g. Landau 1984:227) - planning, writing and producing, which for printed dictionaries follow nicely one after the other. The advent of the computer and the internet have changed the field of lexicography radically, and have affected the way in which dictionaries are constructed and published. Klosa (2013) describes the lexicographical process for online dictionaries (under construction) and identifies six phases (Figure 1).

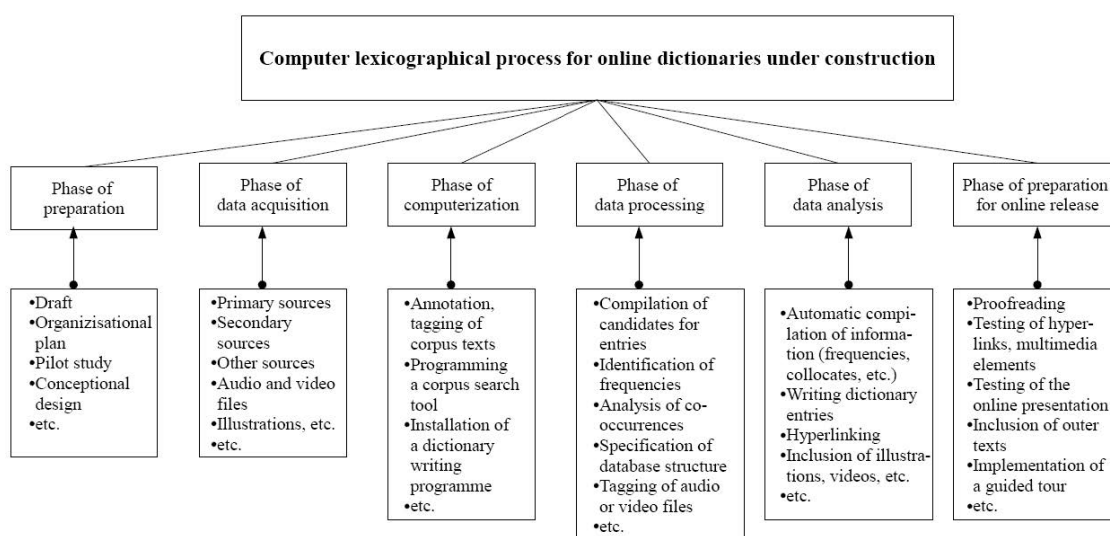


Figure 1 Computer lexicographical process for online dictionaries under construction (Klosa 2013; 520).

For online dictionaries, these six phases do not follow nicely one after the other (e.g. online dictionaries can be published step-by-step: producing an online dictionary may thus begin before the phase of writing is finished) and it is sometimes difficult to identify exact borders between the different phases (Klosa 2013, Tiberius and Schoonheim, To Appear). For a completed dictionary, a 7th phase can be identified, i.e. the afterlife which entails maintenance and preservation of the dictionary after the end of the project. This is similar to what Svensén (2009: 413) calls ‘follow-up’.

The participants at the Bolzano meeting were asked to present the lexicographical workflow for their project using the 6 phases distinguished by Klosa (2013) to find out whether different online dictionaries (under construction) can fit their lexicographical process into the proposed phases.

In addition, the following research questions were raised in the call for papers:

- How to deal with version control and archiving of different versions of the dictionary?
- Print dictionaries mostly follow the alphabetic order of the entries, this is not necessarily so for online dictionaries. What guidelines are used in the different projects for determining which entries should be described and in which order they should be described?
- How can visualisation of lexicographical information influence the lexicographical process?
- How do the changes in the lexicographical process affect the user?
- How can users be involved in the lexicographical process and during which phase should this be done?
- How does automatic extraction of lexicographical data from corpora affect the lexicographical process?

In this deliverable we aim to answer these research questions. First a description of the lexicographical workflow of the different projects represented in WG3 (as of July 2014) is given in Section 2. The order of the project descriptions follows the programme of the meeting. Section 3 addresses user involvement and Section 4 briefly describes a presentation on how visualisation can influence the lexicographical process. Section 5 concludes the deliverable.

2. Lexicographical workflow of WG3 projects

In this section, the lexicographical workflow of the different dictionary projects presented at the Bolzano meeting is described. For each project a short description is given using the same set of properties throughout (cf. Atkins and Rundell 2008:24), followed by a schematic overview of the lexicographical process. Section 2.1 contains the general dictionary projects (6 monolingual dictionaries and 1 bilingual dictionary). Section 2.2 contains the descriptions of the specialised dictionaries and databases (7 projects).

2.1 General monolingual and bilingual dictionaries

2.1.1 [Kielitoimiston sanakirja \(Dictionary of Contemporary Finnish\)](#)

(Institute for the languages of Finland)

URL: <http://www.kielitoimistonsanakirja.fi/>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

language: monolingual (Finnish)
coverage: general; synchronic

| | |
|----------------------------|---|
| size: | 100,000 headwords |
| medium: | print (latest edition 2012); online (latest update 2014) |
| organisation: | word to meaning; sometimes word to word (synonym links) |
| users' language(s): | native Finnish |
| users' skills: | large audience (general public, language professionals, advanced learners) |
| use: | decoding; some encoding |
| update cycle: | in the future, annually |
| team: | 6 lexicographers (4 editors, 1 managing editor, 1 editor in chief) IT resources shared with other dictionary teams |
| project duration: | (with the current name) 2004– |

Overview of the lexicographical process of Kielitoimiston sanakirja:

(update cycle in 4 phases)

Data acquisition: new words and meanings + Evaluation of current version + Feedback on use

Data analysis: adding new and editing existing dictionary entries

Review by managing editor and editor in chief

Preparation for release

2.1.2 [elexiko](http://www.owid.de/wb/elexiko) (IDS Mannheim)

URL: <http://www.owid.de/wb/elexiko/start.html>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

| | |
|----------------------------|--|
| language: | monolingual (German) |
| coverage: | general; synchronic; descriptive |
| corpus: | elexiko corpus; static corpus |
| size: | 300,000 headwords |
| medium: | online |
| organisation: | word to meaning |
| users' language(s): | native German |
| users' skills: | educated adults |
| use: | mainly decoding; some encoding |
| update cycle: | new entries are included directly into the dictionary |
| team: | 5 lexicographers (1 editor in chief, 4 lexicographers) 3 student assistants Connected to the project (but shared with other projects): 1 project manager of the dictionary portal in which elexiko is integrated 2 computational linguists 1 researcher into dictionary use |
| project duration: | ongoing since 2000 |

Overview of the lexicographical process of elexiko:

| Phase | Duration | | | | | | | | | | | | | | | | | |
|-----------------------------------|--------------|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| | Ende 1990 | 2000 | '01 | '02 | '03 | '04 | '05 | '06 | '07 | '08 | '09 | '10 | '11 | '12 | '13 | '14 | '15 | ... |
| Preparation | | | | | | | | | | | | | | | | | | |
| Data acquisition | | | | | | | | | | | | | | | | | | |
| Computerisation | | | | | | | | | | | | | | | | | | |
| Data processing | | | | | | | | | | | | | | | | | | |
| Data analysis | | | | | | | | | | | | | | | | | | |
| Preparation for online release | | | | | | | | | | | | | | | | | | |
| Afterlife | | | | | | | | | | | | | | | | | | |

2.1.3 [Polish Academy of Sciences Great Dictionary of Polish](#)

(*Instytut Języka Polskiego PAN*)

URL: <http://wsjp.pl/>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

- language:** monolingual (Polish)
- coverage:** general; synchronic; descriptive
- corpus:** Polish National Corpus
- size:** currently (end of December 2014 - 21 800 headwords). planned (August 2018) - 50 000
- medium:** online
- organisation:** word to meaning
- users' language(s):** native (Polish)
- users' skills:** large audience (rather than only scholars and linguists)
- use:** mainly decoding; some encoding
- update cycle:** daily
- team:** 40 linguists (scholars and PH.D. students ; (including 20 fulltime); 5 computer programmers (specific-task contracts)
- project duration:** 2005-2018

Overview of the lexicographical process of the Great Dictionary of Polish:

| Phase | Duration | |
|--|-----------|-----------|
| | WSJP 1 | WSJP 2 |
| Preparation | 2005-2007 | 2013 |
| Data acquisition | 2007-2010 | 2013 |
| Elaborating of the computer lexicographic system | 2007-2008 | 2013-2014 |
| Preparing of dictionary entries | 2008-2011 | 2014-2017 |
| Final proofreading of whole dictionary | 2012 | 2018 |

2.1.4 [Slovene Lexical Database](#) (*Communication in Slovene project consortium*)

URL: <http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

language: monolingual (Slovene)
coverage: general; synchronic; descriptive
size: 2.500 headwords (or 10.964 Lexical units: senses, subsenses, multiword expressions and phraseological units)
medium: online: beta version; for download in XML format, including Document Type Definition (DTD) and W3C schema (XSD)
organisation: word to meaning
users' language(s): native (Slovene)
users' skills: general users, school population and partly learners of Slovene as a foreign language
use: mainly decoding; some encoding
update cycle: completed within the project; project has finished
team: 1 project leader; 1 chief editor; 4 lexicographers; 15 beginner lexicographers (students under contract and required training); 0,5 software developer; 1 database manager
project duration: 2008-2012

Overview of the lexicographical process:

| 2008 | 2009 | | 2010 | | 2011 | | 2012 |
|---|---------------------------|----------|---------|----------|---------|----------|------|
| Jun - Dec | Jun | Sept-Dec | Jun | Sept-Dec | Jun | Sept-Dec | Jun |
| Specifications for the continuous collection of written materials for the Reference Corpus of Slovene | | | | | | | |
| Specifications for the compilation of a lexicon of Slovene word forms | | | | | | | |
| Description of the Reference corpus analysis | | | | | | | |
| Specifications for the compilation of an individual lexical database unit | | | | | | | |
| | SLD A-K | | | | | | |
| | Lexicon of word forms A-Ž | | | | | | |
| | | | SLD L-P | | | | |
| | | | | | SLD P-Ž | | |
| Reference Corpus of Slovene with a spoken subcorpus | | | | | | | |

**2.1.5 [Algemeen Nederlands Woordenboek](#)
(Institute of Dutch Lexicology, INL)**

URL: <http://anw.inl.nl>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

- language:** monolingual (Dutch)
- coverage:** general; synchronic; descriptive
- corpus:** ANW corpus; static corpus
- size:** currently 34.702 headwords, target: 70.000 headwords
- medium:** online
- organisation:** word to meaning
- users' language(s):** native (Dutch)
- users' skills:** large audience (ranging from general public to scholars and linguists)
- use:** mainly decoding; some encoding
- update cycle:** every three months a new update is released
- team:** 1 chief editor; 1 project leader; 5 lexicographers; 3 lexicographic assistants; 0,6 computational linguist; 0,5 programmer

project duration: 2000-2018

Overview of the lexicographical process of the ANW:

| | 1 | 5 | 10 | 15 |
|--------------------------------|---|---|----|----|
| Preparation | | | | |
| Data Acquisition | | | | |
| Computerisation | | | | |
| Data Processing | | | | |
| Data Analysis | | | | |
| Preparation for Online Release | | | | |

**2.1.6 [The Dictionary of Contemporary Latvian Language and its Lexicographical Process](#)
(University of Latvia)**

URL: <http://www.tezaurs.lv/mlvv>

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

- language:** monolingual (Latvian)
- coverage:** general; synchronic; descriptive, cultural material, terminology and sublanguages
- corpus:** Balanced Corpus of Modern Latvian
- size:** 45 000 headwords (2017 – 50 000 headwords)
- medium:** online
- organisation:** word to meaning
- users' language(s):** native
- users' skills:** linguists and other language professionals, literate adults, school students
- use:** mainly for decoding
- update cycle:** every 4 years
- team:** 5 lexicographers; 1 editor; 1 computational person
- project duration:** 1997-2017

Overview of the lexicographical process of the Dictionary of Contemporary Latvian Language:

| Phase | Duration |
|------------------------------|-------------|
| Preparation | 1997 - 2001 |
| Writing and Editing | 2001 - 2014 |
| Computerization | 2003 - 2014 |
| Corrections and Improvements | 2014 - 2017 |

**2.1.7 [The Online Dutch-Frisian Dictionary](#)
(Fryske Akademy)**

URL: not known yet

Properties of the dictionary (cf. Atkins and Rundell 2008: 24):

- language:** bilingual (source:Dutch – target:Frisian)
- coverage:** general; synchronic; descriptive
- size:** 70.000 headwords
- medium:** online
- organisation:** word to meaning
- users' language(s):** two specific groups of language-speakers (Dutch and/or Frisian)
- users' skills:** linguists and other language professionals, literate adults, school students, language learners
- use:** mainly for encoding
- update cycle:** every three months a new update is released
- team:** 4 lexicographers (3,5 fte); 1 software engineer (0,4 fte); 1 editor after the end of the project
- project duration:** 2014-2019

Overview of the lexicographical process of the Online Dutch-Frisian Dictionary:

| Phase | Duration |
|--------------------------------|-----------|
| Preparation | 2013-2014 |
| Data acquisition | 2014 |
| Computerisation | 2014 |
| Data processing | 2014-2019 |
| Data analysis | 2014-2019 |
| Preparation for online release | 2014-2019 |
| Afterlife | 2019 |

2.2 Specialised dictionaries & databases

2.2.1 [Valency Database of Croatian verbs](#)

(*Institute of Croatian Language and Linguistics*)

URL: <http://valencijski.ihjj.hr>

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

language: monolingual (Croatian)
coverage: specialised; valency verbs
size: 897 verbs as lemmas
medium: online; basis for online or print dictionary
organisation: word to meaning
users' language(s): native Croatian speakers and learners worldwide of Croatian (two specific groups of language-speakers)
users' skills: linguists and other professionals, language learners
use: encoding (using a word correctly, language teaching)
update cycle: none
team: 7 people (1 project manager and lexicographer; 4 lexicographers; 1 computational linguist and lexicographer; 1 computational linguist and programmer)
project duration: (May) 2013- (September) 2020

Overview of the lexicographical process of the Valency database of Croatian verbs:

| <i>Duration</i> | 2009. | 2010. | 2011. | 2012. | 2013. | 2014. | 2015. | 2016. | 2017. | 2018. | 2019. | 2020. |
|--------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Phase | | | | | | | | | | | | |
| Preparation | | | | | | | | | | | | |
| Data acquisition | | | | | | | | | | | | |
| Computerisation | | | | | | | | | | | | |
| Data processing | | | | | | | | | | | | |
| Data analysis | | | | | | | | | | | | |
| Preparation for online release | | | | | | | | | | | | |
| Afterlife | | | | | | | | | | | | |

2.2.2 [Croatian Metaphor Repository](#)
(Institute of Croatian Language and Linguistics)

URL:

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

- language:** monolingual (Croatian)
- coverage:** specialised; metaphors
- size:** standard
- medium:** web-based
- organisation:** word to meaning (where looking up one word leads to other semantically related words)
- users' language(s):** learners worldwide of the dictionary's language (and all other speakers interested in metaphor)
- users' skills:** linguists and other language professionals; literate adults; language learners
- use:** decoding; encoding
- update cycle:** 1 year
- team:** 10 persons (2 programmers; 1 computational scientist; 3 cognitive linguists; 2 cognitive and computational linguists; 1 cognitive psychologist; 1 cognitive and historical linguist); PI 70%, other collaborators 30% FTE
- project duration:** 3 years

Overview of the lexicographical process of the Croatian Metaphor Repository:

| Month | Objectives |
|-------|--|
| 1-3 | Building Croatian Metaphor Repository framework (M4fw), based on web2py tools |
| 3-36 | Improving and adapting the M4fw to the project needs and results and developing tools for semi-automatic metaphor detection, identifying metaphorical word use and semi-automatic extraction of linguistic metaphors |
| 2-8 | Event structure metaphors |
| 8-12 | Mind metaphors |
| 13 | Seminar: Building Metaphor Repositories: Methods, Risks and Challenges |
| 12-17 | Economics metaphors, governance metaphors, well-being metaphors |
| 17-22 | Time metaphors |
| 22-25 | Morality metaphors |
| 25-31 | Emotion metaphors |
| 31-33 | Cascade metaphors |
| 33-36 | Scalar metaphors |
| 34 | Symposium: Metaphor in Natural Language Processing and Artificial Intelligence |

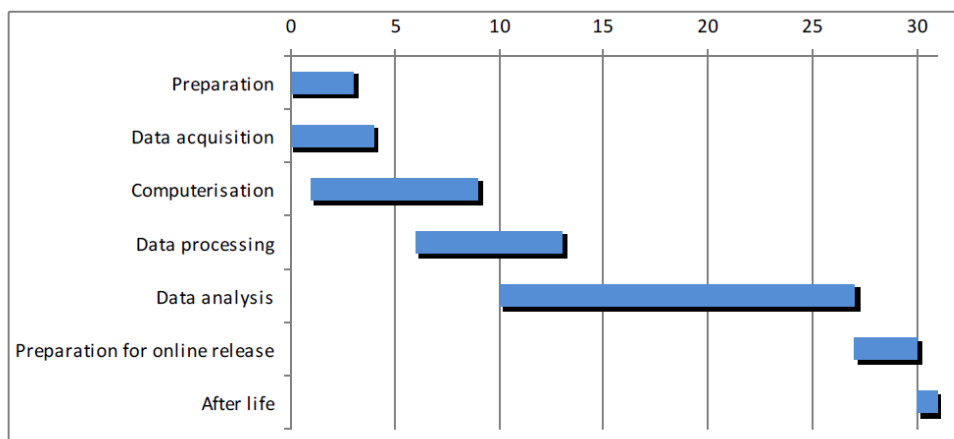
2.2.3 [Portuguese Dictionary of MWE](#)
(Centre of Linguistics of the University of Lisbon)

URL: <http://metashare.metanet4u.eu/repository/browse/lex-mwe-pt-word-combination-in-portuguese-language/8c13600ccd0711e1a404080027e73ea2f9cfd28f51d5437b8f5827c516c348fe/>

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

| | |
|----------------------------|---|
| language: | monolingual (Portuguese) |
| coverage: | specialised; MWE |
| corpus: | 50M balanced Portuguese written corpus |
| size: | 14.000 entries (idiomatic expressions, collocations, institutionalised phrases, favoured co-occurring forms, etc.) + 221.847 manually verified concordances |
| medium: | online |
| organisation: | MWE + concordances |
| users' language(s): | native Portuguese, language learners |
| users' skills: | linguists, computer linguists, language learners |
| use: | encoding |
| update cycle: | project has finished |
| team: | 1 full-time linguist; 3 part-time linguists; 2 part-time computer scientists |
| project duration: | 2004-2006 (30 months) |

Overview of the lexicographical process of the Portuguese dictionary of MWE:



2.2.4 [Estonian Collocations Dictionary](#)
(*Institute of the Estonian Language*)

URL:

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

| | |
|----------------------------|---|
| language: | monolingual (Estonian) |
| coverage: | scholarly; specialised; collocations |
| corpus: | Estonian National Corpus, approx. 560 mln tokens |
| size: | 10,000 headwords |
| medium: | online |
| organisation: | word to meaning |
| users' language(s): | Estonian as a second language or as a foreign language |
| users' skills: | learners at the intermediate and advanced levels (B2 to C1) according to the Common European Framework of Reference for Languages |
| use: | mainly encoding |

update cycle: regular update is not planned, only on neologisms (once a year) and the basis of users' feedback
team: 3 lexicographers
project duration: 2014- 2018

Overview of the lexicographical process of the Estonian Collocations Dictionary:

| Phase | Duration |
|--------------------------------|---|
| Preparation | 2014 |
| Data acquisition | 2014 |
| Computerisation | 2014 |
| Data processing | 2015-2017 |
| Data analysis | 2015-2017 |
| Preparation for online release | on-going |
| Afterlife | Linking the resource with other scholarly dictionaries and language-learning environments |

2.2.5 [Handbook of Slovak Nouns](#)

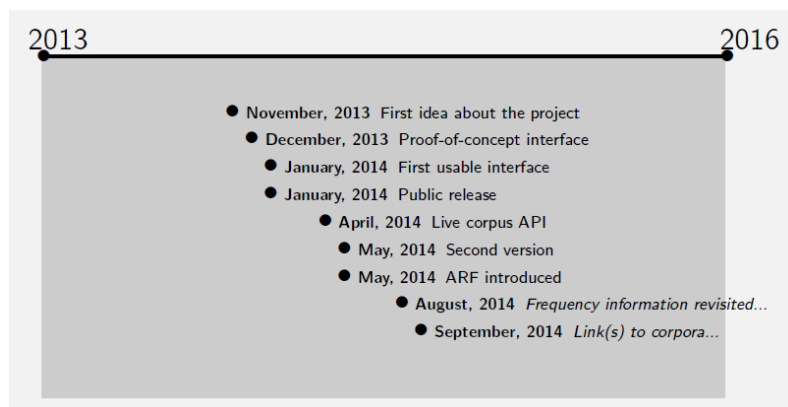
(L. Štúr Institute of Linguistics, Slovak Academy of Sciences)

URL: <http://slovniky.korpus.sk/?d=noundb>;

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

language: monolingual (Slovak)
coverage: specialised; nouns
corpus: korpus.sk
size: 34809 headwords
medium: online
organisation: word to meaning
users' language(s): Slovak (a group of users who all speak the same language)
users' skills: literate adults; school students; language learners
use: mostly encoding
update cycle: permanently under construction
team: 2 persons (morphological database); 1.1 persons (dictionary); 1.5 persons (corpus-technical); 3 persons (corpus acquisition)
project duration: 2013- 2016

Overview of the lexicographical process of the Handbook of Slovak Nouns:



2.2.6 [Portlex](http://portlex.zbr.pt/)

(*Universidad de Santiago de Compostela*)

URL: <http://portlex.zbr.pt/>

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

| | |
|----------------------------|---|
| language: | multilingual (French, Galician, German, Italian, Spanish) |
| coverage: | specialised: NP, valency |
| corpus: | Annotated corpus of the Noun Phrase |
| size: | 200 nouns in Spanish with the entries and equivalents in the others language |
| medium: | online |
| organisation: | word to meaning |
| users' language(s): | learners worldwide of the dictionary's languages |
| users' skills: | language learners, linguists and other language professionals |
| use: | a. decoding: translating from a foreign language text into their own language b. using a word correctly – translating a text in their own language into a foreign language – language teaching |
| update cycle: | |
| team: | 12 researchers plus María José Domínguez Vázquez (all teachers who do this in addition to their normal activities) |
| project duration: | 2013- 2016 |

Overview of the lexicographical process:

| ACTIVITIES | * | FIRST YEAR | SECOND YEAR | THIRD YEAR |
|--------------------------------|---|------------------------------|-------------------|-------------------|
| PREPARATION | | 10 red, 10 yellow | 10 yellow | 10 yellow |
| DATA ACQUISITION | | 10 red, 10 yellow | 10 red, 10 yellow | 10 yellow |
| DATA PROCESSING | | 10 red, 10 yellow | 10 red, 10 yellow | 10 red, 10 yellow |
| DATA ANALYSIS | | 10 yellow, 10 red, 10 yellow | 10 red, 10 yellow | 10 red, 10 yellow |
| COMPUTERIZATION | | 10 red, 10 yellow | 10 yellow | 10 yellow |
| PREPARATION FOR ONLINE RELEASE | | 10 yellow | 10 yellow | 10 red, 10 yellow |

* Before the grant project

2.2.7 [Struna: The Croatian National Termbank](http://struna.ihjj.hr) (Institute of Croatian Language and Linguistics)

URL: <http://struna.ihjj.hr>

Properties of the dictionary/database (cf. Atkins and Rundell 2008: 24):

| | |
|----------------------------|--|
| language: | multilingual (source language: Croatian) |
| coverage: | specialised standardised terminology |
| corpus: | no single corpus used (each project involved in the termbase uses their own resources for compilation) |
| size: | 107.000 terms in 18 professional domains (Struna released in 2012) |
| medium: | web-based only |
| organisation: | concept (representing the meaning) to term, i.e. each meaning is entered as a separate entry |
| users' language(s): | Croatian (a group of users who all speak the same language) |
| users' skills: | translators, literate adults (general public) and high school/university students |
| use: | decoding and encoding |
| update cycle: | released in 2012 |
| team: | 11 people (only 3 permanent positions – 1 project manager, 8 terminologists (linguists, research assistants), 1 language editor, 1 programmer) + every project includes about 10 people (domain experts and 1 or 2 terminologists) |
| project duration: | 16 months per project |

Overview of the lexicographical process:

| Duration (months) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|--------------------------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|
| Preparation | | | | | | | | | | | | | | | | |
| Data acquisition | | | | | | | | | | | | | | | | |
| Computerisation | | | | | | | | | | | | | | | | |
| Data processing | | | | | | | | | | | | | | | | |
| Data analysis | | | | | | | | | | | | | | | | |
| Preparation for Online release | | | | | | | | | | | | | | | | |
| Afterlife | | | | | | | | | | | | | | | | |

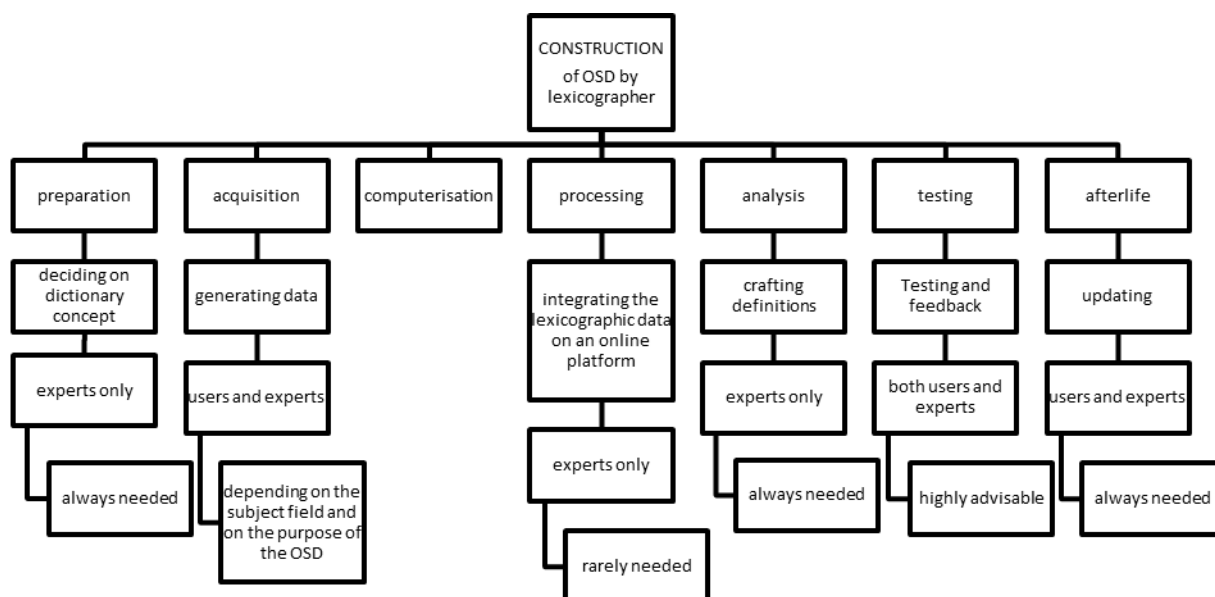
Table 1 Process phases of dictionary project and their time span

3 User involvement

This section deals with user involvement and aims to answer the following research question:

- How can users be involved in the lexicographical process and during which phase should this be done?

Two papers in particular focussed on user involvement. [Patrick Leroyer](#) proposed the following adaptations to Klosa’s schema to reflect user and expert involvement in the construction of online specialised dictionaries:



This diagram shows that in the construction of online specialised dictionaries, there may be user involvement in the phase of data acquisition (depending on the subject field and purpose of the specialised dictionary), there should be user involvement in the phase of testing and that user involvement is definitely needed in the afterlife phase. Experts are involved in more phases during

the construction process. They are potentially involved in 6 of the 7 phases. Only the phase of computerisation does not involve users and experts.

[Stella Markantonatou and Katerina Tzortzi](#) discussed user involvement in a specific project, the Ekfrasis project (a conceptually organised lexicon of Modern Greek). Their project shows that user involvement is important in the phase of data acquisition as corpus and web data only provide the most used structures. However, to also encode less frequent usages, some crowdsourcing method should be devised, given the resource situation for Greek. They used crowdsourcing in their project.

The research question how the changes in the lexicographical process do affect the user, was not discussed at the meeting.

4 *Visualisation*

The presentation by Kris Heylen (University of Leuven) focused on the research question how visualisation of lexicographical information can influence the lexicographical process? He outlined a project entitled „from lexical to semantic sociolectometry“ which although primarily focussing on basic lexicological research will develop tools that can be integrated into the data processing step of the lexicographical workflow.

5 *Concluding remarks*

The WG3 meeting in Bolzano has given us a useful insight into the lexicographical workflow of different projects and allow us to draw some preliminary conclusions about the workflow of online dictionaries. The presentations have shown that overall the different projects can fit their lexicographical process into the phases proposed by Klosa (2013). However, it is sometimes difficult to put exact time labels on the different phases as sometimes a phase continues but does not require full-time effort. This, was, for instance, the case for the ANW dictionary project where the functional design of the web application was carried out at the end of the planning phase extending it by two years. However, writing the functional design was only one of the many tasks carried out during those two years.

The project descriptions also show that even although lexicography becomes more and more computer-assisted, compiling dictionaries still is a highly labour-intensive task. The general monolingual dictionaries of this study have the longest time span with an average of fourteen years. The duration of the compilation of specialised dictionaries/databases is much shorter with an average of just over three years.

Of the different phases the analysis phase takes the longest. This applies to general and specialised dictionaries. In the specialised dictionaries/databases an increase in the data processing phase can be observed (see, in particular, PORTLEX and the Estonian Collocations Dictionary)

The majority of the projects mention the lack of IT support. This is also the case for the more computational projects mentioned under the specialised dictionaries.

The following image from [Domínguez Vázquez et al.](#) (p.5) gives an overview of different phenomena that can affect the initial planning:



During the meeting, attention was also paid to the research question how users can be involved in the lexicographical process and during which phase should this be done. Patrick Leroyer showed in his presentation that in the construction of online specialised dictionaries, there may be user involvement in the phase of data acquisition (depending on the subject field and purpose of the specialised dictionary), there should be user involvement in the phase of testing and that user involvement is definitely needed in the afterlife phase. Stella Markantonatou and Katerina Tzortzi showed the importance of user involvement in the phase of data acquisition for a Greek project. Greek corpora and webdata corpus and web data only provide the most used structures. To also encode less frequent usages, some crowdsourcing is needed. The involvement of the user in general dictionary projects was not discussed and requires further attention.

During the discussion it turned out that some of the research questions that had been raised in the Call for Papers needed a more in-depth study, i.e.

- How to deal with version control and archiving of different versions of the dictionary?
- Print dictionaries mostly follow the alphabetic order of the entries, this is not necessarily so for online dictionaries. What guidelines are used in the different projects for determining which entries should be described and in which order they should be described?
- How do the changes in the lexicographical process affect the user?
- How does automatic extraction of lexicographical data from corpora affect the lexicographical process?

We hope to be able to address them at future meetings. The automatic extraction of lexicographical data will be the topic of the WG3 meeting in August 2015.

References

- Dubois, C. (1990). Considérations générales sur l'organisation du travail lexicographique. In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.), *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*. (HSK 5.2), Berlin/New York, 1574-1588.
- Klosa, A. (2013). The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herberst Ernst (eds.): *Dictionaries. An international Encyclopedia of Lexicography*. Supplement Volume: *Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin, Boston: de Gruyter, S. 517-524. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.4).
- Landau, S. L. (1984). *Dictionaries. The Art and Craft of Lexicography*. New York.
- Svensén, Bo (2009). *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.
- Tiberius, C. and T. Schoonheim (To Appear). The *Algemeen Nederlands Woordenboek (ANW)* and its Lexicographical Process. In: Vera Hildenbrandt (Ed.): *Der lexikografische Prozess bei Internetwörterbüchern*. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik).
- Zgusta, L. (1971). *Manual of Lexicography*. The Hague.