Dictionary of Modern Slovene: lexicographical process

Polona Gantar, Iztok Kosem and Simon Krek

Abstract

This paper describes each phase in the compilation of a database that is to be used as a basis for an online dictionary of modern Slovene and in developing Slovenian language technologies. A proposal for archiving different versions of entries, as well as different versions of the entire database during the compilation process, is also presented. Furthermore, we describe how to include detecting lexical change (the continuous updating of headwords) and dictionary users in the process. This is an important issue in electronic lexicography, but one that still leaves many questions unanswered.

Keywords: lexicographical process, automatic data extraction, online dictionary, detecting lexical change, gradual dictionary compilation

1 INTRODUCTION

The compilation of dictionaries in the digital age is closely linked with modern way of life and access to different types of information via computers or various mobile devices. It is largely driven by the reliability, rapid and free access, and customizability of dictionary content, three characteristics also most valued by dictionary users (Müller-Spitzer et al. 2011). As a result, lexicographers and dictionary publishers are looking for solutions how to provide quality language descriptions with minimum investment of time and money, as well as keep them regularly updated. As leading lexicographers have been pointing out for some time now, it is clear that paper dictionaries, although still present, are becoming obsolete and will eventually no longer be compiled (cf. Krek 2011; Rundell 2014).[1] It is for this reason that planning the compilation of a dictionary is even more important for the language community, especially in Slovenia, where there is currently no corpus-based description of Slovene in existence, and the compilation of the new version of the *Dictionary of Slovene Literary Language* (DSLL2) [2] follows the principles of paper dictionary compilation.

The proposal for the compilation of a dictionary of modern Slovene Language (DMSL: Krek et al. 2013b: 52–60) presents a procedure of dictionary compilation in phases that allows gradual release of dictionary content according to the degree of lexicographic analysis and the amount of information in the entries. The proposal also describes the procedure for regular updating of dictionary entries (ibid.: 46) and the method of prioritising entry treatment (ibid. 45). This paper aims to provide a more detailed description of each phase of the proposed lexicographical process that will meet long-term lexicographic challenges and efficiently utilise all the ICT knowledge and language technologies available, both in terms of methods for extracting language data and ways of presenting dictionary information to users. As the lexicographical processes by which the compilation of digitally-born corpus-based dictionaries are still relatively poorly described,[3] this paper also addresses highly relevant topics such as the inclusion of the language community in the compilation of a dictionary. Furthermore, the problem of the continuous release of dictionary entries is also discussed in this work, including archiving of dictionary information and a developing database version control process.

[1] The future of lexicography was also discussed at a round table titled *Will there still be dictionaries in 2020?*, held at the conference Electronic Lexicography in the 21st Century (eLex, Bled, 10–12 November 2011). A video of this is available at http://videolectures.net/elex2011_bled/.

[2] The dictionary is already being compiled at the Fran Ramovš Institute for the Slovenian Language (FRISL). The dictionary is currently the only general monolingual dictionary receiving government funding, which is problematic because it is based on a concept that focuses on a paper format and thus static dictionary content, and disregards state-of-the-art lexicographic and language technology approaches.

[3] It was for this very reason that the description and planning of the lexicographical process was the topic of one of the workshops of the European Network of e-Lexicography (ENeL) held in July 2014 in Bolzano. The related contributions can be accessed at: http://www.elexicography.eu/working-groups/working-group-3/wg3-meetings/wg3-bolzano-meeting/.

2 PHASES IN DICTIONARY COMPILATION

As a carefully planned process of dictionary compilation, the lexicographical process is one of the key organisational and logistic tasks that affect both the formation and organisation of a lexicographic team, as well as the project timeline and finances. As pointed out by Tiberius and Krek (2014), the existing literature mainly provides descriptions of lexicographical processes in relation to paper dictionary compilation (see Dubois 1990; Landau 1984; Zgusta 1971), consisting of three phases: planning, compilation and publication. Computers (especially the automatic processing of language data), the Internet, and the available quantities of linguistic and related data have undoubtedly affected the way dictionary content is compiled and published. According to Klosa (2013: 4), the lexicographical process in the compilation of non-static online dictionaries consists of six phases, which are not sequential, but can overlap or complement each other (Klosa 2013; Tiberius and Schoonheim 2015). These phases are: preparation, data acquisition, computerization, data processing, data analysis, and preparation for online release.

2.1 The lexicographical process in the proposals for a new dictionary of Slovene

Recently, Slovenian lexicographers have started discussing the need for a dictionary of modern Slovene, but it was not until the publication of the only publicly presented proposal for the compilation of DMSL (Krek et al. 2013b) that such discussions became more concrete. As the lexicographical process is heavily dependent on the dictionary content, methods and main medium for which the dictionary is designed, it represents a key element in the overall concept of a dictionary and how it will be realized.

Before focussing on individual phases in the compilation of DMSL, as proposed by the consortium lead by the Centre for Language Resources and Technologies at the University of Ljubljana,[4] we will present two other related projects: the *New Dictionary of Slovene Literary Language* (NDSLL), the compilation of which is expected to take at least 20 years,[5] and the *Monitor Dictionary of the Slovene Language* (MDSL), which is closely linked with NDSLL and could be seen as a form of dictionary under construction (Klosa 2013: 3),[6] a novelty in Slovenian lexicography.

[4] http://www.cjvt.si/projekti/

[5] See the responses in media to the Proposal of DMSL, e.g. http://www.24ur.com/novice/slovenija/na-nov-slovar-slovenskega-knjiznega-jezika-bomo-cakali-se-leta.html.

[6] Perhaps a more suitable term would be "a never completed dictionary".

2.1.1 Monitor Dictionary of the Slovene Language

MDSL is described as a growing dictionary, and one that is only informative in nature during its compilation.[7] Its initial version contains words which are not found in existing dictionaries of Slovene. but can be found in corpora of Slovene. Also added to the headword list are words that have been unsuccessfully searched for by the users at the website http://bos.zrc-sazu.si, as well as words not found in existing corpora of Slovene, but attested in other, and especially electronic, resources.[8] In the introduction section of MDSL it is also stated that a similar approach will be used for updating the headword list in the future, and that new words will be added to the dictionary every six months. Although the "initial version" is mentioned, the authors do not give any details about how older versions of the dictionary will be archived, or even if is archiving envisaged. The relationship between MDSL and NDSLL is also unclear: "Only time will tell whether individual entries will end up in normative or explanatory dictionaries." (Introduction, MDSL). So despite the ambition indicated by its name, it can be concluded that methodologically, i.e. in terms of the gradual adding of dictionary content, the dictionary does not actually bring any novel lexicographic approach to Slovenian lexicography. The entries are compiled from scratch, and access to different versions is not provided.

2.1.2 The NDSLL concept

An overview of the compilation of NDSLL needs to be made, mainly because the editors claim that the procedure will include three important processes in the pre-editing phase that are nearly identical to the processes envisaged in the compilation of DMSL (Krek et al. 2013b). These are: (a) automatic extraction of data from the corpus, (b) development of a tool for detecting changes in meanings and grammar, and (c) upgrading of existing corpora of Slovene.

The compilation of NDSLL is divided into the pre-editing and editing phases. The pre-editing phase includes the preparation of the headword list, which will serve as a basis for the selection of dictionary entries. The dictionary authors anticipate that corpus data and data from existing dictionaries[9] and other language resources of the Fran Ramovš Institute for the Slovenian Language ZRC

[7] The author and expert consultants claim that the words are selected and described purely from the informative perspective; no normative information is included.

[8] The authors do not provide any details about these resources.

[9] The authors claim that the information from existing dictionaries will be included as much as possible, which casts doubt on their stated intention of compiling a dictionary from scratch (NDSLL: 1, 2).

SAZU[10] will be automatically added to the entries in the dictionary database. Among the listed automatically extracted information are headword spelling, word class, frequency of the lemma and individual word forms, syntactic information, including collocations and examples, certain grammar labels, and certain information on language use, such as treatment of hyphenated words as one word or a multi-word unit. Further analysis of all this information will direct the dictionary treatment of individual headwords. The automatic extraction of the such information from the corpus demands exact decisions about the interpretation of data in terms of the relationships among corpus, lexicon and dictionary, as lemmatisation and morphosyntactic information are closely linked to corpus tagging, which means it is not possible to transfer this information directly into the dictionary entries. The experience from the SLD project shows that this process is by no means trivial. Considering that, at least to the best of our knowledge, these procedures have not yet been tested by the NDSLL team, any evaluation or a detailed presentation of the automatic extraction process cannot be expected. If the authors of NDSLL have decided to use the same methodology for the automatic extraction of data as was applied in the compilation of the SLD, it is then important to stress that the procedure in the SLD project followed very clear methodological guidelines and was adapted to the organisation of dictionary information, which differs in many aspects from the organisation of dictionary information as presented in the NDSLL concept.

The NDSLL concept also envisages the development of a tool that would detect semantic and grammar changes in language use (NDSLL: 78). According to the authors, this would shorten the time of dictionary compilation, as the editors would have the information prepared in advance in the dictionary-writing system (DWS), but in contrast to the purpose of automation (cf. Kosem et al. 2013a) it is anticipated that the editors will check all the information as if they had been designing the entries from scratch (NDSLL: 78).

One of the parts of dictionary compilation is making regular updates to the corpus, a far from a trivial task. There is however no detailed explanation provided on how existing corpora will be updated, how the taxonomy of corpus texts will be upgraded or adapted, how the permission for the new texts will be obtained, and so on, and thus an explanation similar to that provided by Logar (2015) on the updating of the Gigafida corpus is needed. At present, can only find the statement that all the changes made to previous versions of the dictionary will be documented and that, for reference purposes, the users will also have access to older versions of the entries (NDSLL: footnotes 4 and 32).

[10] Based on this list of automatically extracted data we can assume that the team will use the same procedure as was used in the SLD project (Kosem et al. 2013; Kosem et al. 2013a). However, no references are provided in the outline of the NDSLL concept.

2.2 Lexicographic process in the compilation of DMSL

Separate phases in the compilation of DMSL have been first presented in the proposal (Krek et al. 2013b: 52), and in this paper we discuss them in more detail, also according to the experience gained during upgrading of the process of automatic extraction of corpus data and its evaluation (Kosem et al. 2016b). In addition, we consider the experience gained from projects involving the compilation of dictionaries conceived primarily for the online medium and for gradual release, meaning that the dictionary information is available to the users during its preparation. Furthermore, regular updates to completed dictionary entries are anticipated. These are therefore dictionaries which are constantly being compiled and are called *online dictionaries under construction* (Klosa 2013) in the lexicographic literature.

The compilation of DMSL is expected to include five phases (Figure 1), which are designed in a way that enables entry release during the compilation process, i.e. after the first phase. The advantages of this approach outweigh its complexity, shown in the fact that individual phases need to be specified in great detail and the tasks of lexicographers and other team members should be well-defined and coordinated. Multi-phase compilation of entries also enables a more efficient and economical division of work. Most of the work in the first phase is done by a computer, and human input starts in the subsequent phases where lexicographers are used for specific tasks which require lexicographic knowledge and experience. The division of headwords according into the difficulty levels also enables the training of less experienced lexicographers in sense division and definition writing. It is envisaged that certain routine tasks, such as identification and removal of incorrect or irrelevant data and the distribution of collocates and examples under relevant senses, will be left to crowdsourcing (see Fišer et al. 2015), thus reducing the cost of human resources and, most importantly, speeding up the compilation of entries.

During the multi-phase dictionary compilation process, where the availability of different versions of entries is planned, it is of utmost importance that the phase of releasing a dictionary entry is clearly signalled to users. For this reason, we plan to display a different symbol for each phase along with the date of entry release, which reflects the date of the last changes to the dictionary entry. On the one hand, this provides a reference for each entry that the users can use (this is particularly important for researchers and teachers), and on the other it gives some indication to the users as to what they can expect in terms of quantity, treatment and reliability of dictionary content.

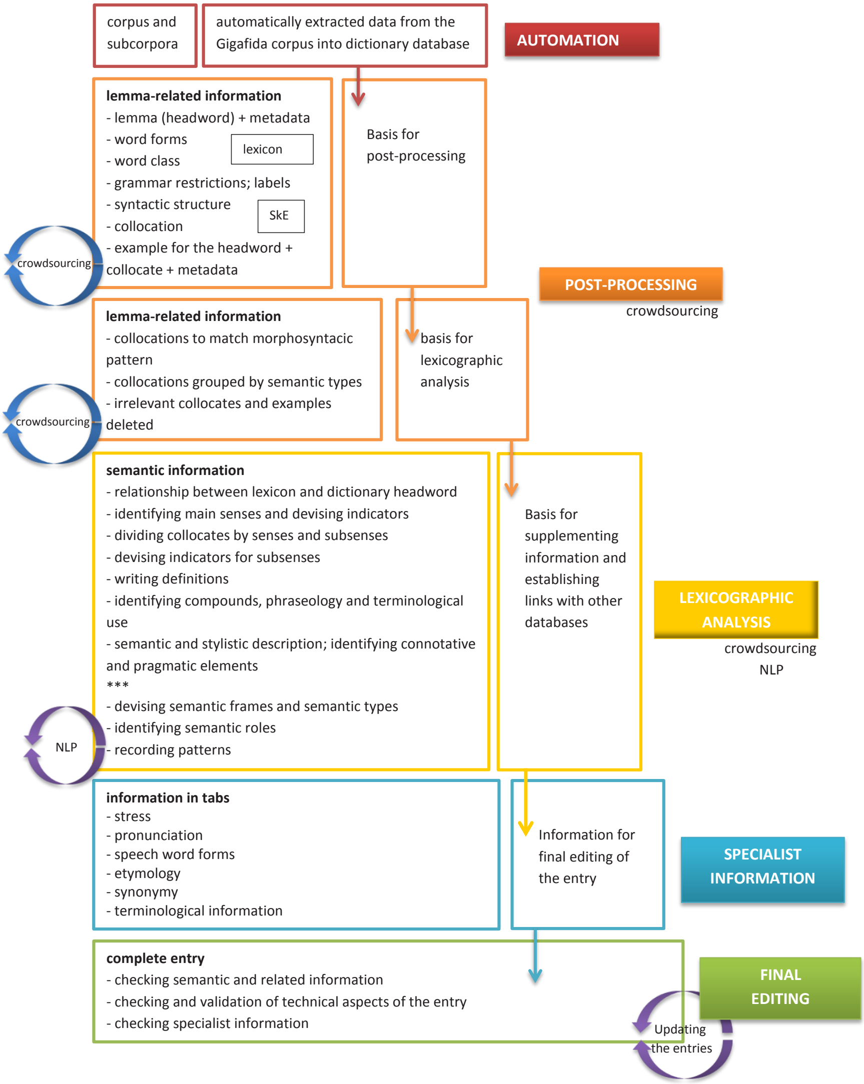

corpus and subcorpora
automatically extracted data from the Gigafida corpus into dictionary database
AUTOMATION
lemma-related information
- lemma (headword) + metadata
- word forms
- word class
lexicon
- grammar restrictions; labels
- syntactic structure
- collocation
SkE
- example for the headword + collocate + metadata
crowdsourcing
Basis for post-processing
POST-PROCESSING
crowdsourcing
lemma-related information
- collocations to match morphosyntactic pattern
- collocations grouped by semantic types
- irrelevant collocates and examples deleted
crowdsourcing
basis for lexicographic analysis
semantic information
- relationship between lexicon and dictionary headword
- identifying main senses and devising indicators
- dividing collocates by senses and subsenses
- devising indicators for subsenses
- writing definitions
- identifying compounds, phraseology and terminological use
- semantic and stylistic description; identifying connotative and pragmatic elements
***
- devising semantic frames and semantic types
- identifying semantic roles
- recording patterns
NLP
Basis for supplementing information and establishing links with other databases
LEXICOGRAPHIC ANALYSIS
crowdsourcing
NLP
information in tabs
- stress
- pronunciation
- speech word forms
- etymology
- synonymy
- terminological information
Information for final editing of the entry
SPECIALIST INFORMATION
complete entry
- checking semantic and related information
- checking and validation of technical aspects of the entry
- checking specialist information
FINAL EDITING
Updating the entries

Figure 1: Phases of the DMSL lexicographic process

2.2.1 Phase 1: automatic extraction

Phase 1 in entry compilation consists of automatic extraction of lexical data from the Gigafida corpus (Logar Berginc et al. 2012), which will also be used when devising the headword list. In addition to using the frequency information from Gigafida, the headword list will, based on various statistical calculations, utilise information from the Kres corpus, Gos corpus (Verdonik and Zwitter 2011) and other freely available corpora of Slovene. In order to obtain a more specialised vocabulary, the compilation of specialised corpora is envisaged, as well as updates to existing corpora, e.g. the compilation of a subcorpus of textbooks (see Logar 2015; Vintar and Logar 2015 for more). Furthermore, we plan to use thematic tagging of domain texts in the form of corpus metatags, which are then transferred into the dictionary database during automatic extraction (see Gantar and Kosem 2013; Kosem 2015).

Taking into account the structure of a dictionary entry in DMSL (see Klemenc et al. 2015), the automatically extracted data are as follows:

- **Lemma** in the basic form, as found in the Gigafida corpus and the Sloleks lexicon of Slovene word forms (Dobrovoljc et al. 2015), and all of its word forms (offered in a separate tab).
- Corpus or sub corpus **frequency** of the lemma.
- Word class, based on the word form tag in Gigafida and Sloleks.
- Certain **grammatical alerts** related to typical syntactic or contextual behaviour of the lemma in the corpus, such as frequent use with proper names, predominant use in third person or when citing (verbs). This information is extracted from the corpus using a combination of the directives CONSTRUCTION and UNARY in the Sketch Engine tool (Kilgarriff et al. 2004) and is presented in the dictionary database in the form of alerts, which can be later (in Phase 3) converted into dictionary labels such as *pogosto zanikano* ('often in negative'), *pogosto v 3. os. ednine* ('often in 3rd person singular') etc. (see Kosem 2015).
- **Syntactic structures,** identified during manual analysis of word sketches for the purposes of SLD and used as a basis for a new, improved version of sketch grammar for Slovene (Krek 2012a).

This new sketch grammar utilizes the directives *CONSTRUCTION, *COLLOC and *SEPARATEPAGE. The first enables the identification of grammatical relations without collocates, which is particularly useful for extraction of corpus examples containing all elements in verb patterns, such as "subject-predicate-indirect_object-direct_object", confirming the existence of the pattern for the particular verbal headword. The second directive is used to identify elements that are categorized as syntactic combinations in the lexical database, such as the statistically significant "preposition-noun-preposition" combinations. The third directive is intended for creating a separate word sketch page for relations with three elements (directive *TRINARY), which enables the introduction of relations with prepositions that can have more specific definitions: for example, they produce a separate word sketch for each noun case (of the six cases in Slovene) in "noun-preposition-verb" or "noun-preposition-noun" patterns. This new sketch grammar for Slovene thus provides a very fine-grained overview of a word's collocational behaviour and is devised solely for automatic extraction of lexical data. The word sketches produced by such a sketch grammar are difficult to process by a human user, due to the high number of relations and their complex naming system.

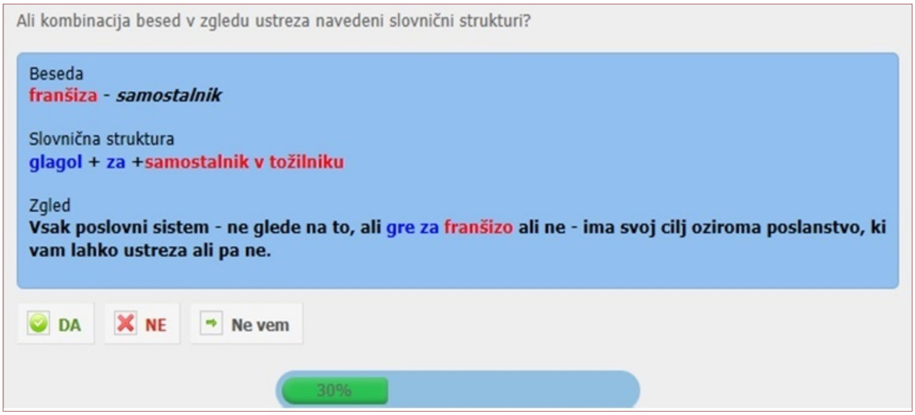
- **Collocates** found with the lemma in a particular syntactic structure and forming potential collocations, syntactic combinations and compounds. The latter are identified and recorded under the relevant sense by lexicographers in Phase 3.
- **Corpus sentences** containing the lemma and collocate in a particular syntactic structure. Corpus sentences are extracted with the GDEX tool (Kilgarriff et al. 2008), with configurations based on the GDEX for Slovene (Kosem et al. 2011) but especially adapted for automatic extraction (Kosem et al. 2013b). The extracted sentences are candidates for inclusion in the dictionary (they may require minor modifications), and are thus potential dictionary examples.

The automatic extraction procedure has already been tested in the compilation of the SLD (Kosem idr. 2013, 2013a), where an API script using different parameters for each grammatical relation was used to automatically extract the above listed types of data, which were then imported into the dictionary database in the iLex DWS (Erlandsen 2004). We also conducted an evaluation of the procedure, comparing it with the manual entry compilation (Kosem et al. 2015), and later upgraded and improved the procedure for the use with the *Collocation Dictionary of the Slovene Language* (Gantar et al. 2015). Among the most notable upgrades are automatic removal of collocates that have all the same examples, and automatically converting the lemma and/or the collocation in the word form with appropriate case, gender and number according to the syntactic structure. In addition, the initial values used for extraction were improved considering the frequency and word class of the lemma (see Kosem et al. 2013a; 2013b for more), resulting in the development of several parameter configurations for each word class. In the improved procedure we also extracted collocates using salience and frequency order, respectively, and then combined both sets of data. This enables us to select the most relevant collocates for each lemma.

2.2.2 Phase 2: post-processing and clean-up

Phase 2 is intended for (a) post-processing, which includes (semi-)automatic removal of errors and irrelevant data, also by using crowdsourcing, and (b) adding of metatags which enable the connecting of information in the dictionary database and establishing links with other dictionary databases (e.g. the initial dictionary database, collocations dictionary database, and synonym dictionary database). Automatically extracted data can be additionally improved with post-processing, e.g. by putting collocates in the appropriate gender and case according to the syntactic structure, and by forming collocation sets which include semantically related collocates. In order to facilitate combining the information from different databases, it is necessary to tag different elements within collocations (e.g. prepositions, conjunctions, and reflexive pronouns of verbs) and/or add relevant linking information in the tag attributes of the lemma or its collocates (e.g. ID from Sloleks).

We envisage the use of crowdsourcing to remove irrelevant collocations, which are the consequence of errors in lemmatisation or are simply corpus noise. Figure 2 shows an example of a task in which the crowdsourcers are asked to decide whether the combination in the automatically extracted sentence (coloured in blue and red in the example; *gre za franšizo*) reflects the identified syntactic structure:



Ali kombinacija besed v zgledu ustreza navedeni slovnični strukturi?

Beseda
franšiza - ***samostalnik***

Slovnična struktura
glagol + za +samostalnik v tožilniku

Zgled
Vsak poslovni sistem - ne glede na to, ali gre za franšizo ali ne - ima svoj cilj oziroma poslanstvo, ki vam lahko ustreza ali pa ne.

Figure 2: The crowdsourcing task for removing irrelevant collocations and examples in the SLD.

The crowdsourcing tasks were conducted with the SlowCrowd tool (Tavčar et al. 2012), which has also proved useful in the improvement of the Slovene version of wordnet called SloWNet (Fišer 2009). The initial tests during the compilation of the SLD (Kosem et al. 2013b) have shown that the use of crowdsourcing for data clean-up is reliable, and can considerably reduce the time spent on this phase of the lexicographic process.

2.2.3 Phase 3: lexicographic analysis

The next phase involves the lexicographic analysis of data, which makes it most demanding in terms of expertise and logistics, and it also takes the most time. Tasks include sense division, definition writing, identification of grammatical, syntactic, normative and stylistic characteristics of words and their meanings.

In this phase, the lexicographers are presented with cleaned-up automatically extracted data for each lemma. Word class information is automatically attributed to the lemma, and is the same as the morphosyntactic tag in the Sloleks lexicon. Consequently, lexicographers' first task is to check the correspondence between the lexicon unit and the dictionary entry. This is by no means easy, and the efficiency and congruity of lexicographers' decisions relies on providing them with detailed instructions containing all possible situations and common solutions, especially in terms of homonymy and conversion, i.e. in accordance with the decisions outlined in the dictionary concept (see Gantar 2015 and Dobrovoljc 2015). In DMSL, there is a symmetric relationship between the entry in the lexicon and in the dictionary, while potential exceptions are signalled in the database using a predetermined set of machine-readable restrictions.

The main tasks of lexicographers in this phase are identifying senses and subsenses, and writing definitions for priority entries. The entry structure follows that of the entries in SLD, which means that the lexicographers also devise sense indicators that represent a constituent part of a sense menu, which offers an overview of the entry senses and subsenses, and, at certain noun and adjective entries, they must also devise semantic frames that contain a typical valency pattern of a particular (sub)sense. Also important in this phase is adding information intended for natural language processing, for example sentence patterns, semantic types (similar to the approach used in Corpus Pattern Analysis; Hanks 2004; Hanks and Pustejovsky 2005) and semantic roles. The lexicographers also identify and write definitions for compounds – marking those that require an input from the terminologists – and phraseological units.

Lexicographic work is organised according to the difficulty level of the entry and the availability of templates for semantically related entries. To achieve the optimal efficiency, the tasks are divided between (a) experienced lexicographers who perform sense division and write definitions, identify more complex grammatical and syntactic patterns, and record any stylistic and pragmatic information about the word's usage; (b) lexicographers specialised in phraseology, description grammatical and syntactic characteristics of individual (sub)senses, and normative information; and (c) relatively inexperienced lexicographers who conduct less demanding lexicographic tasks, such as checking whether collocates have been correctly assigned to senses and syntactic combinations during crowdsourcing, forming collocation sets, and identifying context for compounds and phraseological units.

Lexicographic tasks that can be regarded as routine in nature and do not require considerable lexicographic knowledge will be left to crowdsourcing. One such task is assigning automatically extracted corpus sentences (and the collocation in a particular syntactic structure they attest) to one of the (sub)senses.[11] The secondary goal of the task is to get feedback on the suitability of the sense division and to identify unidentified or new senses.

At the end of Phase 3 the majority of relevant semantic information is already available to the users. The next step consists of adding information which will be presented in the online dictionary separately (e.g. in tabs), and this is done in Phase 4.

2.2.4 Phase 4: adding specialist language information

Phase 4 of the lexicographic process comprises of adding information from other databases and enriching existing entry information. This phase requires the involvement of experts from other fields, especially terminologists, and linguists with expertise in standardisation and norms. The following information is added to the entries at this point:

- information on **spelling**, based on the Sloleks lexicon and according to the connection between the lexicon entry and the dictionary entry. This includes detecting the overlap, or lack of it, between the pronunciation and declension paradigms of the headword, which is one of the criteria for detecting homonymous lemmas;
- information on **pronunciation** based on the Gos corpus and predetermined procedures, including marking the stress, pronunciation of word forms and providing pronunciation that cannot be deducted from the headword's spelling (Jurgec 2015);
- information on speech word forms and any special semantic characteristics observed in the speech corpus Gos (Verdonik 2015);
- information on **etymology**, more specifically on the origin of the word and its related word forms in different languages, and information about archaic forms of the word in and the time period in which they appeared;
- information on **synonyms**, obtained using the Sketch Diff feature in the Sketch Engine tool and information from SloWNet, and

[11] This task is described in more detail in Fišer et al. (2015).

- **terminological analysis of data.** For this task, we will form a network of experts in different fields and develop an online platform that will facilitate the monitoring and coordination of work.

2.2.5 *Phase 5: final editing*

The final phase in the compilation of DMSL is intended for final editing of the entry and a consistency check of the information found in different tabs. Here the lexicographer's task is to check the consistency of information with the dictionary concept and real language use, and to check the validity of entry structure. The lexicographer has the option to edit, expand or even return the entry to one of the previous phases if they, for example, identify inconsistencies in sense division, or in compound or phraseology treatment, or find incomplete terminological information.

Another important step of this phase is automatic detection of semantic changes in a word's usage which can, if identified, return the entry to Phase 3 (sense division, multi-word unit identification, crowdsourcing, and so on), thus requiring another final editing at a later stage. So, even though Phase 5 represents the end of lexicographic process, entry compilation again involves automatic extraction, in this case of relevant new data, found in updated corpora.

3 UPDATING THE DICTIONARY DATABASE

The dictionary database plays a very important role in the process of entry compilation and their eventual presentation to users, representing the source of all dictionary information on one hand, and the archive of all the decisions made during the five phases of the lexicographic process on the other. Considering that the phases of dictionary compilation are clearly delineated, the dictionary database needs to include a workflow that can at any time provide the editors and lexicographers with the information on the phase status of each entry. Another level of complexity in planning the dictionary database is introduced by two connected decisions: regular updating of the dictionary and the option of uploading the entries after each phase is completed.

By regular updating of the dictionary we mean updates to the existing entries in the database that have already gone through all the phases of the lexicographic process, as well as the compilation of completely new entries, especially priority ones (e.g. neologisms). The latter need to have in the dictionary database a special warning about their importance, distinguishing them from other entries, which in turn provides the editors with the option to alert users to such entries once the dictionary is updated. This is also true for updates to already completed entries, which can be made in the form of adding new information (e.g. senses, collocations, and phraseological units) based on the analysis of new data (e.g. monitor corpus or new version of the reference corpus) or in the form of modifying existing information (e.g. correcting errors). In this case, more important for the users is the temporal aspect, i.e. when was the new information added (and based on which resource).

A special case in regular updating of the dictionary is replacing old with new information, but only in terms of presentation to the users. For example, at some point the decision could be made to replace existing examples with new ones (cf. Klein and Geyken 2010; Lemnitzer et al. 2015). To enable this, examples (and other microstructural elements) in the database need to include the information on whether they are part of the online dictionary entry or not. This makes it possible to keep all the examples in the database while showing the users only those that are most relevant.

Releasing entries after each phase does not require any additional information, except that related to the workflow; existing dictionary entries that are being updated with new information should be excluded from this procedure, as the combination of analysed and non-analysed data could confuse users. More relevant for such a procedure are visualisation solutions which also require certain types of information (e.g. date of release and version).

It is therefore essential to prepare a procedure that gives the lexicographers a clear idea about the phase of the entry, date of its inclusion in the online dictionary, and the date of addition of new information (the completion date of the new version[12]). We believe that such a procedure can ensure an efficient and transparent lexicographical process, and facilitate clear and understandable presentation of dictionary content to users.

4 TREATMENT OF DIFFERENT VERSIONS AND PRESENTATION

This section is dedicated to three questions relevant for updating the dictionary using the proposed lexicographic process: How often should the updates be

[12] It is important to distinguish between entry versions denoting larger changes (e.g. after the completion of each phase or after updating existing dictionary entries with new information), and entry versions in the DWS. Namely, the DWS records every single change made to the entry, thus enabling the comparison of two database "versions", reviewing and restoring of deleted data, etc. It is thus recommended to consistently use terminology that distinguishes between the two processes.

made? How to clearly distinguish between incomplete entries or incomplete entry information from the completed ones? How to handle different versions of dictionary entries and different versions of the dictionary?

Dictionaries of other languages use two different approaches to updating online dictionaries: updates in regular intervals (usually every few months) or continuous updates, as soon as new entries are compiled. The former approach is used by dictionaries such as the *Oxford English Dictionary* (OED),[13] which releases updates every four months and also has a special webpage dedicated to promoting the updated entries. A similar approach is used by the *Macmillan English Dictionary* (MED),[14] although instead of separately providing the information on the date of the updates,[15] the MED alerts the users to selected new entries in the New Words section on the front page.

The latter approach, i.e. immediate release of completed entries, is used by the *Comprehensive Dictionary of Polish*[16] (Žmigrodzki 2014) and the *Dictionary of Contemporary Dutch*[17] (Tiberius and Schoonheim 2015). It is noteworthy that these two dictionaries are being made from scratch, and are thus real dictionaries under construction. Consequently, the motivation for continuous release of dictionary content is much higher, in terms of satisfying both the users and funders, than at dictionaries that merely add new words or update existing entries. In view of this, the proposal in the NDSLL concept to update the dictionary annually (NDSLL: 3) is not ambitious enough, and fails to fully consider needs of the users. As such, we would expect that the users, who have been waiting for a new description of Slovene for over 25 years, will be offered the results of lexicographic work as soon as possible.

The approach of immediate release is also part of the lexicographic process in the proposed DMSL, where it is envisaged that entries will be released after each phase of their compilation. There are already dictionaries in Slovenia using this approach, such as iSlovar,[18] a dictionary of computer terms, which distinguishes between four phases of entry compilation: “predlog” (‘proposal’; proposed by the editor or user), “pregledano” (‘reviewed’; reviewed by the editor), “strokovno pregledano” (‘reviewed by expert’; reviewed and edited by experts) and “urejeno” (‘edited’; reviewed by the dictionary team; this is the final editing). It should be pointed out that the updates are not made every day or even every hour, but are made as packages (i.e. several entries at the same time) in frequent intervals, resulting in greater transparency for both lexicographers and users.

[13] http://www.oed.com/

[14] http://www.macmillandictionary.com/

[15] The website with FAQ (http://www.macmillandictionary.com/faq.html provides the information that the dictionary is updated several times a year.

[16] Wielki słownik języka polskiego: http://www.wsjp.pl/

[17] Algemeen Nederlands Woordenboek: http://anw.inl.nl/show?page=search1.

[18] http://www.islovar.org

The alerts about the entry status serve to distinguish between incomplete and complete entries. The proposal for the compilation of DMSL (Krek et al. 2013b: 52–60) suggests that the entry status is indicated with coloured dots (from red to green) and the date of last update (Krek et al. 2013b: 27). The date can be used to distinguish between different versions of the entry at the same phase (e.g. when updating the completed dictionary entry). The final visualisation solution may end up being different from the one in the proposal, but will need to include at least these two types of information. In addition, there will be, like on the OED website, a webpage dedicated to the updates, offering a list of new and updated entries and their status, as well as information on any major changes.

One of the important decisions related to the immediate release of entries concerns the treatment of previous versions. This is somewhat less problematic as far as releasing entries at different phases is concerned, as the version most relevant for the users is the one that contains the largest amount of information. This was in fact one of the hot topics at the OED Symposium in 2014, as a few participants complained about that after updates they could no longer see the previous versions of entries. Their argument was that by updating the definitions we lose information on how a certain sense or usage of the word was perceived by lexicographers working on the entry at a particular point in the past. While this argument is perfectly legitimate, it is worth pointing out that the OED is a historical dictionary in which a diachronic view of language use is of vital importance.

The decision on whether to include the option of comparing different versions of the dictionary entry in DMSL will be based on the findings of surveys among dictionary users. Nonetheless, access to previous versions of entries is already envisaged for the researchers working in the fields of linguistics, machine learning and natural language processing. Namely, we plan to make freely available new versions of the dictionary database, which will be released simultaneously with updates to the online dictionary, except in cases when the changes made will be relevant only to the dictionary database. An important part of this process will be detailed documentation, which will include not only a description of changes to the dictionary entries, but also a description of all the content and technical changes to the dictionary database, e.g. new types of database labels, changes in DTD (*Document Type Definition*) and so on.

5 CONCLUSION

The lexicographic process of dictionary compilation that envisages the continuous release of entries at different phases of compilation, regular updates and access to different versions of the entries, is a complex procedure, demanding a well-designed and detailed strategy which affects the organisation of lexicographic work. The lexicographic process proposed in this paper is based on the automatic extraction of lexical data, which are in the subsequent phases first cleaned of incorrect and irrelevant data, and then analysed and supplemented with additional information. A distinction should be made between the information in the dictionary database which is part of the regular work on the entries and is not presented to users, and the database information in different versions of the entries offered to users. It is also vital that lexicographical process is clearly defined and recorded, enabling the lexicographers and editors to carry out continuous and consistent dictionary work. This also makes it possible to provide the users with information on the status of the entry, the entry information that is available at each phase, and with different versions of the entries, the latter being particularly relevant for the purposes of research, further processing or teaching. It is important to note that the proposed lexicographic process is devised for an online dictionary, using a specific entry structure and internal organisation of information, and containing different types of lexico-grammatical information linked both within the entry (e.g. sense menu, collocations, syntactic structures, patterns, examples under senses, collocations, compounds and phraseological units) and with the information in other tabs (word forms, speech, norm, synonyms etc.).