Slovene Lexical Database: lexicographical process

Polona Gantar, Iztok Kosem, Simon Krek
Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia,
Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia,
Jožef Stefan Institute, Ljubljana, Slovenia
E-mail: apolonija.gantar@guest.arnes.si, iztok.kosem@trojina.si, simon.krek@guest.arnes.si

1 Introduction

Lexicographic process of compiling the Slovene Lexical Database (SLD) is closely linked with two main purposes for which the database was designed (see below); at the same time, the SLD needs to be considered in the context of the current status of dictionaries of Slovene, namely a lack of dictionaries describing contemporary Slovene. The only monolingual dictionary of Slovene is the Dictionary of Standard Slovene (SSKJ), which was conceived in 1960s, with the last volume published in 1991. The dictionary was conceived for print and was compiled entirely using lexicographic slips, with information based mainly on 19th century Slovene fiction texts. In 2012, the Dictionary of newer standard Slovene words (SNB) was published, but was unable to fill the gap in the language description of Slovene between 1991 and 2012, partly due to its conceptual links with SSKJ and the lack of corpus-based methodology in its concept. These circumstances have been very important for designing individual phases in the process of compiling the SLD for two reasons:

a) the process needed to include a detailed overview of state-of-the-art in dictionary making, lexicographic methodology, corpus tools etc. and the implementation of these approaches and methods to the characteristics of the Slovene language;
b) the Slovene language needed a completely new language description due to social and political changes, and the changes brought by digitization and rapid development of information and communications technologies; all these circumstances have also contributed to semantic changes of the lexis already found in the SSKJ.

2 Slovene lexical database (SLD)

The Slovene Lexical Database (SLD) is one of the results of the Communication in Slovene project[1], and has two main objectives: firstly, to provide a framework for a dictionary of contemporary Slovene that includes a) descriptions of relevant types of lexico-grammatical information, b) a description of methods for extracting such information from corpora, including the development and adaptation of corpus tools according to special features of the Slovene language, c) a description of lexicographic procedures, based on corpus lexicography, and d) a description of the way the data is organized in the database. Secondly, parts of lexico-grammatical information are enhanced with formalizations to be used by natural language processing tools and language technologies for Slovene.

At the project's conclusion, the SLD contained 2,500 content-word entries or 10,945 lexical units, i.e. senses, subsenses, multi-word units and phraseological units. The database is conceptualized as a network of interrelated lexico-grammatical information (sense, syntax, collocations, examples), with lemma, or the headword, representing the top hierarchical level and functioning as the umbrella for all lexical units placed under it.

It should be pointed out that the SLD was not conceptualised with a specific dictionary in mind, but instead represents a sort of theoretical concept, a pilot study for an online dictionary aimed at digital natives. For this reason, the entire database was published online at the end of the project as a demo version of the Online dictionary of Slovene, and was also the point of departure for the Proposal for a dictionary of contemporary Slovene (Krek et al. 2013).

3 Lexicographical process of the SLD

The starting point for determining the phases of the SLD lexicographic process, which are presented in Table 1 in the Appendix, was Klosa's (2013; see also Tiberius and Schoonheim in print) division of the lexicographic process for an online dictionary under construction. Changes to the names of the phases and their internal division are due to the fact that we were compiling a dictionary database rather than a dictionary and that the SLD was part of a project with pre-determined schedule and contents. One of the findings when developing the SLD was that the phases, which can be found in any dictionary project – planning, writing and producing (Landau 1984: 227) – are overlapping when one is

[1] The operation was partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. Project web page: http://eng.slovenscina.eu/.

making a dynamic online dictionary. At the same time, the lexicographic process is different as computers and language technologies play a more pivotal role, and an important role is also given to experts in website design.

3.1 Phase of preparation

Considering the specifics associated with the inception of the SLD we divided the phase of preparation into four sub-phases: a) organizational phase that includes the plan for human resources, finances, and the project schedule; b) an overview of existing lexicographic projects; c) a preparation of concept that includes the main conceptual characteristics of the dictionary or the dictionary database, and d) a preparation of plan for the implementation of language technologies and computers which according to Klosa's phases (2013: 520) includes the phase of data acquisition, the phase of computerization and the phase of data processing.

3.1.1 SLD timeline

At the beginning of every dictionary project, a detailed organizational plan that include a number of people involved, their tasks, and the financial plan of the project needs to be prepared. In the case of the SLD, this was already included in the project documentation, so we will not discussed it further here. The SLD was closely connected with other activities in the Communication in Slovene project. In Table 2, which shows the project timeline, only the two most directly linked activities are pointed out, namely the compilation of a reference corpus and a lexicon of word forms.

<table>
<tr><th>2008</th><th colspan="2">2009</th><th colspan="2">2010</th><th colspan="2">2011</th><th>2012</th></tr>
<tr><th>Jun - Dec</th><th>Jun</th><th>Sept-Dec</th><th>Jun</th><th>Sept-Dec</th><th>Jun</th><th>Sept-Dec</th><th>Jun</th></tr>
<tr><td>Specifications for the continuous collection of written materials for the Reference Corpus of Slovene</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>Specifications for the compilation of a lexicon of Slovene word forms</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2">Description of the Reference corpus analysis</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2">Specifications for the compilation of an individual lexical database unit</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2"></td><td colspan="2">SLD A-K</td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td colspan="4">Lexicon of word forms A-Ž</td><td></td><td></td></tr>
<tr><td colspan="4"></td><td>SLD L-P</td><td></td><td></td></tr>
<tr><td colspan="6"></td><td colspan="2">SLD P-Ž</td></tr>
<tr><td colspan="8">Reference Corpus of Slovene with a spoken subcorpus</td></tr>
</table>

Table 2: Timeline for the SLD and other activities of the Communication in Slovene project, related to the lexicographic process

The SLD-related activities were thus taking place within the deadlines set for the project and the phases are described below. Dictionary treatment of lemmas did not follow the originally envisaged division according to alphabetical order (see Table 1); lemmas were divided into three equal segments, and headwords for lexicographic analysis were selected from a larger headword list based on the type of lexicographic problem.

3.1.2 Overview of current lexicographic projects

The first part of the preparation phase (June 2008 – December 2008) was dedicated to a) a detailed analysis of European dictionary projects and language technology projects, with particular attention being paid to new corpus-based lexicographical methods, b) evaluation of corpus tools and dictionary-writing systems (IDM, Tlex, iLex ...), c) a description of electronic lexical databases and electronic and online dictionaries, lexico-grammatical information in them and their (online) presentation. Based on these analyses, we selected examples of good practice and individual segments of lexicographic description that were later used in the conceptualisation of the SLD.

3.1.3 Conceptualisation of the SLD

The second part of the preparation phase (January 2009 – June 2009) involved the development of the concept for the SLD and description of the procedure for analysing the reference corpus. At the same time as we were developing the SLD concept, which included describing types of lexico-grammatical information and their organization in the lexical

database, the activity was taking place of determining parameters for collecting texts to increase the existing reference corpus (FidaPLUS) into a new 1,2-billion-word corpus called Gigafida (Logar et al. 2013). The preparation of a corpus for lexical analysis also required the selection of a tagging system, compilation of a training corpus, and development of a tagger and a parser for Slovene. At the end of the conceptualisation phase, two project deliverables were produced: Description of the Reference Corpus analysis (Gantar et al. 2009) and Specifications for the compilation of an individual lexical database unit (Gantar et al. 2009a). The deliverables presented the basis for the SLD Style Guide for the lexicographers, which was regularly updated throughout the SLD project.

3.1.4 Preparation of language technologies

While the SLD concept was being prepared and before we could start compiling database entries, we had to organize the computational or language technological aspect of the SLD project. This phase included the development of a platform for file storage and data exchange between lexicographers (intranet), selection of one of the existing corpus tools (Sketch Engine) and developing the necessary algorithms for Slovene, including sketch grammar (Krek and Kilgarriff 2006) and GDEX configurations (Kosem et al. 2011), uploading the corpus and localization, and the selection and installation of one of the dictionary-writing systems. During the conceptualisation of the SLD and manual compilation of entries, when a larger number of lexicographers were working on the project, we used Entry Editor, a program that is part of a larger dictionary-writing system called DPS (IDM). Later, during the automation phase we replaced Entry Editor with iLex (Erlandsen 2004).

This phase also included the preparation of a headword list which was limited to 5.000 most frequent lemmas in the corpus. When developing the concept, we decided to select only content words. Lemma selection was also influenced by the aim to cover as many lexicographically problematic topics as possible, e.g. homonymy, spelling variants, conversion etc. Lemmas were divided into different difficulty groups to enable an optimal evaluation of lexicographic work. The need for specific technical improvements (the development of internal programs for corpus data analysis) and upgrade of corpus tools (writing different version of sketch grammar, updating the list of syntactic structures, changes to DTD etc.) was confirmed by the evaluation of sample entries and later throughout the compilation of the SLD; consequently, these activities were not limited to a particular phase in the lexicographic process.

3.2 Data analysis (lexicographic analysis and data processing)

The largest amount of time in the lexicographical process was dedicated to the compilation of entries (July 2009 – September 2011). This phase was also the most demanding in terms of content. A team of lexicographers included four experienced lexicographers, who were fully employed during the project, and 15 beginner lexicographers, who were under contract and required training. Initially, entries were selected from the headword list according to the lexicographic problem that we wanted to systematically solve and describe in the Style Guide. For each lemma, a corpus sample of 150-300 concordances was prepared (depending on the corpus frequency of the lemma) and manually analysed in order to detect senses and subsenses, multi-word units and phraseological units, pragmatic information etc. The analysis of the selected concordances also helped identify typical patterns and any usage information, which was then registered with labels, e.g. if a word was typically used in a particular register, text type, domain etc. Analysing sample concordances, the lexicographer created a draft sense division of the lemma, and then confirmed the existing information, and added new information, with the analysis of the word sketch for the lemma in the Sketch Engine. Word sketches were also used as a source of collocations and related syntactic structures and corpus examples. An important part of the entry compilation stage was continuous evaluation of the quality of the data, provided by the word sketch in comparison with the results of manual analysis of the word's concordances. Based on this analysis, an updated version of sketch grammar (Krek 2012) was developed in the second stage of entry compilation, and the Slovene configurations for the tool for automatic identification of good corpus examples were also prepared (Kosem et al. 2011)

An important aspect of the SLD was the development of a procedure for automatic extraction of syntactic structures and related collocations and examples from the corpus, and their direct transfer into dictionary-writing system. This phase was possible only at the end of the SLD project, when all relevant syntactic structures had been registered and the database DTD was finalised. The preparation of the automation procedure included: a) the selection of lemmas for data extraction; b) the development of sketch grammar, specifically designed for automatic extraction; c) GDEX configuration, or configurations, for automatic extraction; d) the preparation of API script for data extraction via Word Sketch in the Sketch Engine; and e) specifying the parameters such as minimum collocation frequency, minimum salience of a grammatical relation, and minimum salience of a collocation. Using automatic data extraction, we compiled approximately 150 entries; the time of compilation was reduced by approximately 25 %, if compared with manual analysis.[2]

[2] It should be pointed out that many headwords were less frequent in the corpus. We would expect greater time savings at more frequent words.

3.3 Preparation for online release

The SLD can be downloaded in XML format,[3] as initially envisaged in the project documentation. The decision to publish the database online was made after the project, with the intention of demonstrating how the data in the lexical database can be transferred into an online dictionary, which enables multi-layered presentation of data and various search and visualization options. Audio and video content, as well as links to other dictionary resources, were not part of this project; however, the links to corpus concordances are provided.

The preparation of the SLD for online publication included decisions on how to visualize different types of information, and prioritizing the data according to their relevance. For example, we developed multi-layered presentation for the types of data that are mainly of interest for lexical analyses (e.g. typical patterns for each sense) and natural language processing (e.g. syntactic structures). Also, we needed to select the most appropriate type(s) of searching the database. As we had a relatively small number of entries, we decided to use a combination of a search window and lists of entries for each letter of the alphabet.

The transfer of the SLD into an online form was not overly ambitious due to lack of human resources and funding, but it nonetheless showed the importance of cooperation between lexicographers, computer experts and web designers. At the same time, the process pointed out that the visualisation of different types of dictionary information has a considerable impact on how that information is organized in the dictionary database. For example, collocations were sorted according to the relevance of syntactic structures, which had to be taken into consideration at automatic extraction of collocations from the corpus.

3.4 Afterlife

After the conclusion of the project we published the Proposal for a dictionary of contemporary Slovene (Krek et al. 2013). The document includes a detailed description of procedures for making an online dictionary on the basis of examples of good lexicographic practice, language technology procedures and corpus methodology. The proposal is based on the SLD concept of registering and structuring lexico-grammatical data, and offers a detailed description of the envisaged project in terms of lexicographic analysis, technical support, organization and finances. The proposed dictionary would offer different types of information on the Slovene language: semantic information, grammar information, notes on norm, etymology and multimedia content. During the dictionary-making process, priority would be given to topical lexis, and the dictionary would be updated on a regular basis as a part of a carefully planned process that includes automatic detection of new words and meanings, lexicographic analysis (including using crowd-sourcing for cleaning automatically extracted data), and immediate display of lexicographic entries online.

The publication of the Proposal proved very important for plans on making new dictionaries of Slovene, as the Proposal served as the basis for Consultation on a new dictionary of Slovene[4], an event organized by the Ministry of Culture of the Republic of Slovenia. In addition, a consortium lead by the Centre for Language Resources and Technologies of the University of Ljubljana[5] was established recently, and its first project will be the compilation of a dictionary based on the concept described in the Proposal.

4 Conclusion

This paper presented the lexicographic process of the Slovene Lexical Database, so the paper describes the compilation of a dictionary database rather than a dictionary. The SLD project is interesting for state-of-the-art lexicographic practice due to its focus on an online dictionary as its final outcome, in terms of both dictionary concept and data organization. The lexicographic process described here needs to be put into the context of current Slovene dictionary situation where no contemporary dictionaries for Slovene are available; as a result, the lexicographic process had to include the analysis of current situation in corpus lexicography and language technologies, and training of lexicographers in corpus analysis. Division of phases of the SLD compilation also showed that the compilation of an online dictionary, which assumes regular updating of dictionary content, demands a close collaboration between lexicographers, computational linguists, IT staff, language technologists, and web designers. This collaboration should already begin at the conceptualisation phase, but is especially important in the phase of preparation which in the SLD's case included organizational, methodological-theoretical, conceptual and language technology stages. It is also clear that a great deal of energy, in terms of both the financial aspect and human resources aspect, should be dedicated to automation of lexicographic procedures and regular updating of contents. In this respect, the online medium offers significantly more options for

[3] http://eng.slovenscina.eu/spletni-slovar/prenos.

[4] http://www.mk.gov.si/si/delovna_podrocja/sluzba_za_slovenski_jezik/predstavitev_podrocja/dogodki/posvet_o_novem_slovarju_slovenskega_jezika/.

[5] http://cjvt.si/.

presenting dictionary contents as the print format, but at the same time it presents new challenges to lexicographers and computational experts.

5 References

Erlandsen, J. (2004). iLex – new DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004). Brno, 6. – 7. september 2004*. Available at: http://nlp.fi.muni.cz/dws2004/pres/#15.

Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) *Natural language processing, multilinguality: sixth international conference, Modra, Slovaška, 20-21 Oktober 2011*, pp. 72-80.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) *Proceedings of the 11th Euralex International Congress*. Lorient: Universite de Bretagne-Sud, pp. 105-116.

Klosa, A. (forthcoming): 'New developments in lexicographic theory IV: Research in dictionary production and use' 26. The lexicographical process (with special focus on online dictionaries) in: *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography.* Edited by Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, Herbert Ernst Wiegand. Berlin/New York: de Gruyter.

Kosem, I., Gantar, P., Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In Kosem, Iztok et al. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper: proceedings of eLex 2013 Conference, 17-19 October 2013*, Tallinn, Estonia. Ljubljana: Trojina, Institute for Applied Slovene Studies; Tallinn: Eesti Keele Instituut, 2013, pp. 32-48.

Kosem, I., Husák, M., McCarthy, D. (2011). GDEX for Slovene. In I. Kosem, K. Kosem (eds.) *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011, Bled, 10-12 November 2011*. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 151-159.

Krek, S. (2012). New Slovene sketch grammar for automatic extraction of lexical data. Presented at *SKEW3 workshop*, 21-22 March 2012, Brno, Czech Republic. Available at: http://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw

Krek, S., Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec, J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference*. Ljubljana, Slovenia.

Krek, S., Kosem, I. Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*, v1.1. Available at: http://trojina.org/slovar-predlog/datoteke/Predlog_SSSJ_v1.1.pdf

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Landau, S. L. (1984): Dictionaries. The Art and Craft of Lexicography. New York.

Lemberg, I. (2001): Aspekte der Online-Lexikographie fur wissenschaftliche Worterbucher. In: Lemberg, I./Schroder, B./Storrer, A. (eds.), Chancen und Perspektiven computergestutzter Lexikographie. Tubingen, 71&91.

Tiberius C., Schoonheim T. (2014). The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process.Preprint to appear in Vera Hildenbrandt (Eds.) *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie"*. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/2014).

Table 1: phases in the SLD lexicographical process