UDK 811.163.6’373
811.163.6’374
Pregledni rad
Rukopis primljen 23. XII. 2019.
Prihvaćen za tisak 18. III. 2020.
doi.org/10.31724/rihjj.46.2.7

Polona Gantar
Faculty of Arts, University of Ljubljana
Aškerčeva cesta 2, SI-1000 Ljubljana
apolonija.gantar@ff.uni-lj.si

DICTIONARY OF MODERN SLOVENE: FROM SLOVENE LEXICAL DATABASE TO DIGITAL DICTIONARY DATABASE

The ability to process language data has become fundamental to the development of technologies in various areas of human life in the digital world. The development of digitally readable linguistic resources, methods, and tools is, therefore, also a key challenge for the contemporary Slovene language. This challenge has been recognized in the Slovene language community both at the professional and state level and has been the subject of many activities over the past ten years, which will be presented in this paper.

The idea of a comprehensive dictionary database covering all levels of linguistic description in modern Slovene, from the morphological and lexical levels to the syntactic level, has already formulated within the framework of the European Social Fund’s Communication in Slovene (2008-2013) project; the Slovene Lexical Database was also created within the framework of this project. Two goals were pursued in designing the Slovene Lexical Database (SLD): creating linguistic descriptions of Slovene intended for human users that would also be useful for the machine processing of Slovene. Ever since the construction of the first Slovene corpus, it has become evident that there is a need for a description of modern Slovene based on real language data, and that it is necessary to understand the needs of language users to create useful language reference works. It also became apparent that only the digital medium enables the comprehensiveness of language description and that the design of the database must be adapted to it from the start. Also, the description must follow best practices as closely as possible in terms of formats and international standards, as this enables the inclusion of Slovene into a wider network of resources, such as Open Linked Data, BabelNet and ELEXIS. Due to time pressures and trends in lexicography, procedures to automate the extraction of linguistic data from corpora and the inclusion of crowdsourcing into the lexicographic process were taken into consideration.

Following the essential idea of creating an all-inclusive digital dictionary database for Slovene, a few independent databases have been created over the past two years: the

Collocations Dictionary of Modern Slovene, and the automatically generated Thesaurus of Modern Slovene, both of which also exist as independent online dictionary portals. One of the novelties that we put forward together with both dictionaries is the 'responsive dictionary' concept, which includes crowdsourcing methods. Ultimately, the Digital Dictionary Database provides all (other) levels of linguistic description: the morphological level with the Sloleks database upgrade, the phraseological level with the construction of a multi-word expressions lexicon, and the syntactic level with the formalization of Slovene verb valency patterns. Each of these databases contains its specific language data that will ultimately be included in the comprehensive Slovene Digital Dictionary Database, which will represent basic linguistic descriptions of Slovene both for the human and machine user.

1. Introduction

The following paper describes the compiling of the *Dictionary of Modern Slovene*, an on-going project the beginnings of which date back to 2008. It examines the development of its database design, the crucial decisions that shaped it, and lessons learned. It also presents its most notable results, focusing on individual online dictionaries and databases completed so far and those currently in development. The paper concludes by outlining the current state of affairs, i.e. the design of the primary objective of the project, the *Digital Dictionary Database*, and lays out plans for the future.

2. Long-Standing Issues

Ever since the 1990s, the Slovene language community was in need of a comprehensive language description, which would reflect the current state of the Slovene language by taking into account language changes and state-of-the-art lexicographic methods of production. These were some of the key recurring issues:

– existing language data was not accessible to the entire research community;
– there was no description of modern Slovene;
– there was no unified concept for standardization on the levels of Grammar–Dictionary–Orthography;
– there was no model of constant vocabulary monitoring.

3. State-of-the-Art Lexicography

Keeping in mind the above issues and based on examples of good lexicographic practice, we determined the key guidelines for the compiling of the *Dictionary of Modern Slovene.*

The dictionary was envisioned as a comprehensive description of Slovene in one place. It, therefore, needed to contain information regarding sense, multi-word expressions, grammatical information, language use, normative alerts and constraints, details about pronunciation, etc. Furthermore, it had to be available online, preferably through a unified web interface, which would enable dictionary database searching and links to outside language resources related to specific language areas. This new description of Slovene was to be based on real data, primarily obtained from various Slovene corpora – general as well as specialized – and other existing or developing digital language resources. The language description had to be reliable, and the information clear and easily obtainable.

Dictionary data would be retrieved from a digital dictionary database and formalized in a manner that would enable the creation of various dictionaries and at the same time, serve as a useful platform for multiple NLP tasks. The automatization procedures were to form an essential part of the lexicographic process. We were looking for an efficient trade-off between digital processing, which necessarily includes errors and manual, time-intensive and costly analysis. The extraction and structuring of language information would be automatized, and the lexicographic and language analysis tools improved on the basis of thorough linguistic evaluation. These findings would be used to improve the processes of automatization, up until the level where machine processing could no longer substitute for human interpretation.

From the very outset, the dictionary was envisioned as a born-digital dictionary. Its online presentation would fully employ all the advantages of web features in the visualization and interlinking of language data. One of the critical elements in dictionary design is knowing the users, which is why we dedicated considerable effort to user analysis. The process of compiling a dictionary benefits greatly from the inclusion of the language community. Various crowdsourcing tasks were thus integrated into the individual phases of the lexicographic process, in the stages of cleaning, structuring as well as evaluating of corpus data.

Another vital aspect of the design of a born-digital dictionary is its sustainability: the dictionary should track language changes, such as semantic shifts, variants, trends, and norms. The design of the dictionary database also needs to consider documentation methods and ways of archiving different time variants. Individual phases of the lexicographic process should allow for the addition and editing of language data in the dictionary database.

Finally, we envisioned an open-access dictionary: the dictionary database and the online dictionary were designed to be accessible under the Creative Commons Licence.

We also wanted our digital dictionary and its comprehensive language description to take part in other European initiatives related to common dictionary infrastructure and interoperability of dictionary content. Such initiatives have already taken place within the framework of the European ENeL COST Action[1] and are currently being undertaken mostly within the scope of project Elexis.[2]

4. Digital Dictionary of Modern Slovene: Project History

The idea of an online dictionary of modern Slovene with all the features mentioned above has evolved gradually, over some time. The bulk of the necessary linguistic infrastructure was developed within the scope of the *Communication in Slovene* project, which took place from 2008-2013 and was financed by the European Union Social Funds. One of the key results of the project was the *Proposal for the Dictionary of Modern Slovene*, which was put forward to the lay and professional public in 2013.[3] The theoretical background regarding the lexicographic features of the dictionary was described in the monograph *Lexicographic Description of Slovene in the Digital Environment* (Gantar 2015), which defined the structure of the lexical database and the automatization procedures to be employed in the lexicographic process. Further theoretical and practical aspects of the key dictionary segments were presented in the monograph *Dictionary of Modern Slovene: Problems and Solutions* (Gorjanc et al. 2017).

[1] www.cost.eu/actions/IS1305/#tabs|Name:overview.
[2] https://elex.is.
[3] Predlog za izdelavo Slovarja sodobnega slovenskega jezika: www.sssj.si.

4.1. Communication in Slovene

The *Communication in Slovene*[4] project was an important predecessor of the *Dictionary of Modern Slovene*. During its course, several corpora were developed: the written corpus Gigafida, with 1.2 billion words, the balanced corpus KRES, spoken corpus Gos, learners' corpus Šolar, and the training corpus SSJ500k, which is manually annotated on the levels of tokenization, lemmatization, morphosyntactic tagging, name entity recognition, parsing, multi-word units and semantic roles. The project also developed the initial tools for processing corpus information, such as the syntactic parser and morphosyntactic tagger. The obtained corpus data served as the basis for the first language descriptions of modern Slovene, namely the *Slovene Lexical Database* and *Sloleks, the Slovene Morphological Lexicon*. The project also resulted in the creation of web portals, such as the *Orthography Guide*, *Pedagogical Grammar*, and *Corpus Concordancers.*

4.1.1. Slovene Lexical Database

The creation of the *Slovene Lexical Database* (SLD) was vital for all further steps in the compiling of the digital *Dictionary of Modern Slovene*. Its main purpose was to create a starting point for the lexicographic description of modern Slovene, which could serve as a template for the compiling of digital dictionaries and provide lexico-grammatical information in a form suitable for digital processing and development of language technology applications for Slovene. The SLD is also essential for the type and manner of organization of linguistic data in the database. As shown in Figure 1, language data was organized as a network of interconnected lexico-grammatical information on six hierarchical levels. Crucially, all the linguistic data is subordinate to or organized in relation to the headword sense.

[4] http://eng.slovenscina.eu.

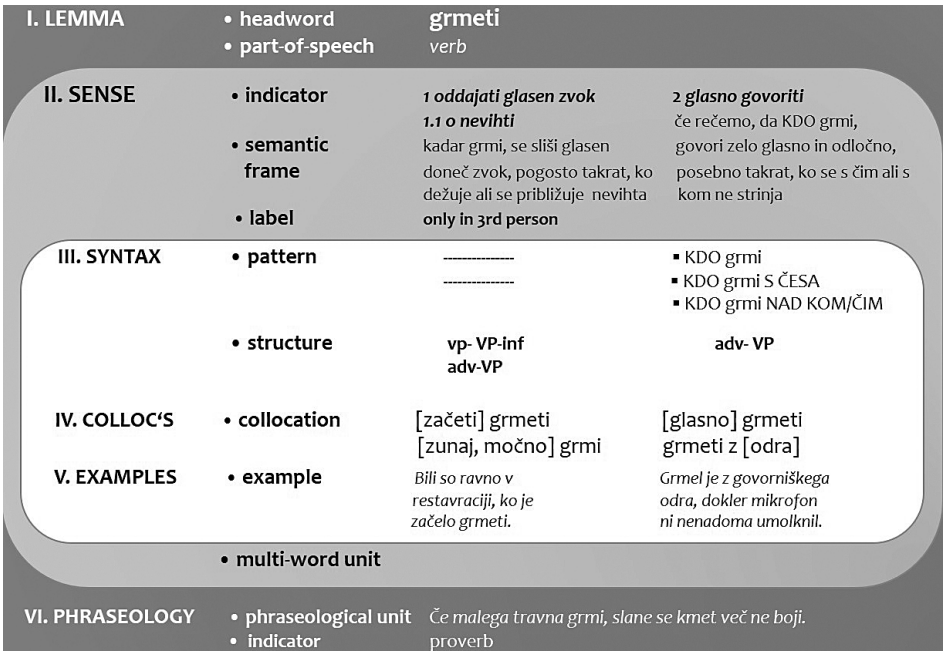

Figure 1: Structure of the *Slovene Lexical Database*

The first level gives the basic form of the headword, including part of speech information. Following is the most important, semantic level, which describes the senses and sub-senses with two types of information: semantic indicators and semantic frames. Conceptually speaking, semantic frames are akin to frames in the FrameNet project and to what P. Hanks refers to as the "prototypical syntagmatic patterns" in the CPA system.[5] In verbs, as well as some nouns and adjectives, semantic frames record the argument structure and semantic types found in a particular sense and sub-sense. Semantic frames thus establish a link between a particular sense of the headword and the syntactic conditions necessary for its realization. The third level describes syntactic information; it formalizes syntactic patterns for the purpose of digital processing, more precisely, for the automatic extraction of collocations from the corpus. The fourth level presents the patterns and structures verified by typical collocates. The fifth level presents the corpus examples, which verify the lexical and grammatical information presented on the previous levels. At the time of developing the SLD, we obtained

[5] https://nlp.fi.muni.cz/projects/cpa.

examples by using the GDEX function in the Sketch engine tool (Kosem et al. 2011).

4.1.2. Automatization in the Compiling of Dictionary

During the building of the lexical database, we tested several procedures for the automatic extraction of data from corpora and the possibility of data transfer to specific locations in the dictionary database (Gantar et al. 2016). The process involved the automatic extraction and export of specific data – such as headwords, part of speech, grammatical information in the form of labels, syntactic structures, and collocations with corpus examples – directly into the dictionary writing system. The subsequent evaluation established that a semi-automatic approach reduces by half the amount of time needed to create an average-sized dictionary entry.

By the time we completed the SLD, we had determined the basic steps of the lexicographic process that makes possible the monitoring of language changes, continuous enrichment of dictionary entries, and their regular updating. The basic stages of the lexicographic process thus include automatic extraction, removal of errors trough post-processing and crowdsourcing, lexicographic analysis of data, the addition of specialized language information – in particular, terminological description, stress, pronunciation, etymological information, etc. – and the final editing and coordination of information for the entry as a whole. We also examined ways of involving the language community in the lexicographic process through various crowdsourcing tasks (cf. Fišer et al. 2015).

5. Individual Digital Dictionaries and Databases

As already mentioned, the comprehensive digital dictionary of modern Slovene has evolved in stages. After presenting the key dictionary components in the monograph *Dictionary of Modern Slovene: Problems and Solutions* (Gorjanc et al. 2017), we started to work on compiling individual databases and online dictionaries. Currently, there are three open-access digital dictionaries and dictionary databases available online: *Thesaurus of Modern Slovene*, the *Collocations*

Dictionary of Modern Slovene, and the *Slovene Morphological Lexicon*. As part of continuous updating, we upgraded the basic reference corpus Gigafida, creating the Corpus of Standard Modern Slovene Gigafida 2, and developed a new interface, primarily aimed at the general public.

The design of each dictionary is predicated on its content; what they all have in common, though, is the idea of a unified web interface, the automatic extraction, and processing of data in the first phase, the inclusion of the language community into the process of their compiling and the provision of open access to the research community.

5.1. Thesaurus of Modern Slovene

The Thesaurus is based on *The Oxford DZS Comprehensive English-Slovenian Dictionary*, the collocation database, and the Gigafida corpus; it currently contains approximately 105,000 headwords and 350,000 synonyms. The Thesaurus was created by applying the methodology of automatic identification, extraction, and organization of synonym candidates (Krek et al. 2017). This facilitates the regular updating and upgrading of the dictionary database. Apart from that, the Thesaurus enables the users to compare the use of different synonyms in collocations, through the application of Sketch Differences, a Sketch Engine function, and by linking to examples in the Gigafida corpus. This type of dictionary is best described as a responsive dictionary since it includes the language community into the lexicographic process by click-enabling it to add new synonyms and clean the noise caused by automatic processing (Čibej and Arhar 2019). Typical issues related to user involvement in dictionary compiling usually concern user motivation and competency. To boost motivation, the interface was designed to offer various elements typically found in crowdsourcing projects, for example, a list of users with most added synonyms, a list of synonyms with the largest number of upvotes, "task of the day", etc. User activity is regularly tracked through the analysis of user logs; user suggestions are checked and incorporated into regular updates. Individual user activity can be reconstructed, which means that poor or potentially trolled entries can be identified and removed.

5.2. Collocations Dictionary of Modern Slovene

The primary aim of the *Collocations Dictionary of Modern Slovene* is to present as much information as possible about Slovene collocations, on the basis of modern material and with the use of the latest computer and lexicographic methods (Kosem et al. 2018). The current version of the dictionary contains around 36,000 headwords and more than 7 million collocations. The dictionary is intended for native speakers as well as learners of Slovene as a foreign language. The compiling of the dictionary follows a set of predetermined stages, which indicate the level of lexicographically processed data. The first stage involves work with automatically extracted corpus information, which is then further processed or post-processed. The following phases include the removal of inadequate information, for example, syntactic structures and collocations which need further improvement in terms of reliability. This is followed by matching collocations to appropriate senses, in the case of polysemous headwords. The last phase involves the final editing of the entry and includes the revision and removal of poor or inadequate examples. In the future, entries will be updated depending on the availability of relevant sources. The first of these updates will be based on data from the new Gigafida corpus. The *Collocations Dictionary* can also be considered a responsive dictionary, the two main criteria being that the data is acquired automatically and that it includes user involvement.

5.3. Slovene Morphological Lexicon

Sloleks, the Slovene Morphological Lexicon is an upgrade of the morphological dictionary conceived during the *Communication in Slovene* project (Dobrovoljc et al. 2017). The Lexicon contains basic information on Slovene words, particularly their word class, inflected forms, and grammatical characteristics. The second version of the Lexicon includes around 100,000 headwords and more than 2.5 million word forms. The Morphological Lexicon has two primary aims: to provide a lexicon for use in language technology applications and to support the creation of normative manuals on orthography and pronunciation. All information contained in the Lexicon corresponds to corpus information, which means that information in the Lexicon can be directly linked to labels in the Gigafida

corpus. The second version of the Lexicon also includes automatically generated word stress, pronunciation recordings and IPA phonetic transcriptions of all the word forms. Following the example of responsive dictionaries, such as the Thesaurus and the *Collocations Dictionary*, the morphological dictionary allows the user to evaluate automatic stress and add user-made pronunciation recordings.

6. Digital Dictionary Database

The creation of the *Digital Dictionary Database* (3D) has been one of our main goals ever since setting out to compile the *Dictionary of Modern Slovene* 10 years ago. The main idea is to have a single, central database of Slovene, which can be enriched with different types of language information and from which data can be exported to each individual dictionary within the comprehensive *Dictionary of Modern Slovene*. In order to achieve this, we designed a central relational database, where all the language data and data extracted from corpora are stored in a pre-defined data model (Figure 2).

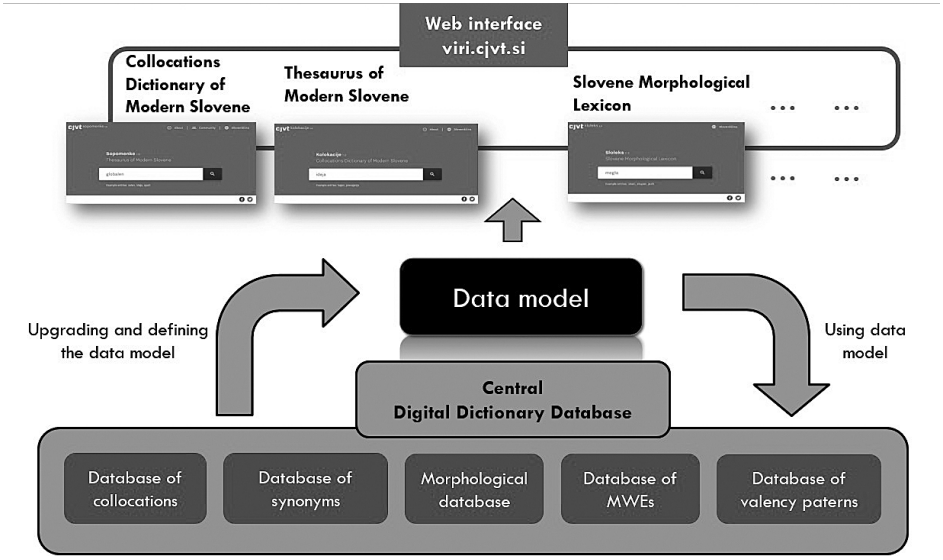

Figure 2: Structure of the Digital Dictionary Database

The central database contains several s.c. satellite databases that employ the same data model. The 3D currently contains data for the *Collocations Diction-*

ary and the *Morphological Lexicon*. In the future, however, it will also include several other databases. The specifics of each of these databases will, in turn, determine the organization of data in the data model.

The central element of the model is a lexical unit (LU), which needs to have at least one sense; each sense may include one or more types of definitions. There are various types of lexical units – a word, a multi-word unit, a collocation, etc. – and they can all have a different status, according to the individual phase of the lexicographic process. Lexemes form lexical units. Each lexeme includes several types of morphosyntactic data – word class and its features, word form, etc. – from the Morphological Lexicon Database. In multi-word lexical units, each LU is determined by syntactic structure. Syntactic structure information includes structural elements (or "components") and their relations, i.e. labels defined by treebank systems. Corpus sentences are linked with LU senses and may be labelled as examples that explain a specific LU or collocation.

At the moment, the data model offers technical solutions for core types of data – representation of various types of lexical units, sense structure, collocations, syntactic patterns, examples, etc. – but does not include representation of multilingual data.

6.1. Future Developments

In the future, the data model will be adapted on a semantic level, with the addition of different types of semantic information. We wish to add different types of sense descriptions, such as semantic indicators and whole-sentence definitions. One of the greatest challenges ahead, however, and one of the main issues that need to be addressed on an interlingual level is linking to universal sense concepts. A considerable amount of integration will be necessary in the coming years, including work on common formats, lemma lists, as well as cross-linking references from dictionaries to corpora.

7. Conclusion

A comprehensive database covering all segments of linguistic description includes data that can be accessed by users – in the form of digital dictionaries, language applications, and online textbooks – as well as used for machine processing, especially for machine learning and development of semantic language technologies. As such, the idea is not new in the European context. Among others, current examples of such work include the Unified Data Model for Presenting Lexical Data EKILEX, which is being developed at the Institute of the Estonian Language (Tavast et al. 2018), and the joint project of the Centre for Digital Lexicography of the German Language, which aims to provide a ubiquitous search interface for diverse dictionary sources (Geyken 2019).

The Centre for Language Resources and Technologies at the University of Ljubljana aims to develop a comprehensive dictionary database, i.e. the basis for the digital *Dictionary of Modern Slovene*, through the continuous building of processable sources, such as corpora, lexical databases and lexicons, and the development of methodologies for automated gathering and processing of language data. The specific characteristics of individual dictionaries demand a complex system of stable or core database elements; at the same time, the database should be flexible enough to allow optimal upgrading, documenting, and archiving of language data. The transmission of data between the database and dictionary interfaces should be automated and should enable an entire range of lexicographic editing. As already mentioned above, one of our main priorities remains to take part in initiatives promoting compatibility and interoperability of language data within open-access frameworks, such as multilingual WordNet, BabelNet, CILI (Collaborative Interlingual Index) and Linguistic Linked Open Data (LLOD).

Acknowledgment

Over the past ten years, a number of researchers have worked towards developing our idea as a whole. They have been affiliated with institutions including the Faculty of Arts, the Faculty of Computer and Information Science and the Jožef Stefan Institute, all of which operate within the framework of the Centre for Language Resources and Technology of the University of Ljubljana. Our work has been funded by the European Social Funds and the Slovene Ministry for

Education, as part of project *Communication in the Slovene language*; several of our projects were supported by the Slovenian Research Agency – *New Grammar of Modern Slovene* (J6-8256) and *Collocations as a Basis for Language Description: Semantic and Temporal Perspectives* (J6-8255) – and carried out within the framework of programmes *Language Resources and Technologies for Slovene* (P6-0411) and *The Slovene Language – Basic, Contrastive and Applied Studies* (P6-0215).

References

Čibej, Jaka; Arhar Holdt, Špela. 2019. A crowdsourcing cleanup of the thesaurus of modern Slovene. *Electronic lexicography in the 21st century: proceedings of eLex 2019 Conference*. Eds. Kosem, Iztok; Krek, Simon. Lexical Computing. Sintra. 338–356. https://elex.link/elex2019/wp-content/uploads/2019/10/eLex-2019_Proceedings.pdf (accessed 31 March 2020).

Dobrovoljc, Kaja; Krek, Simon; Erjavec, Tomaž. 2017. The Sloleks morphological lexicon and its future development. *Dictionary of modern Slovene: problems and solutions*. Eds. Gorjanc, Vojko et al. Ljubljana University Press. Ljubljana.

Fišer, Darja; Čibej, Jaka; Dobrovoljc, Kaja; Gantar, Polona; Kosem, Iztok; Arhar Holdt, Špela; Popič, Damjan; Erjavec, Tomaž. 2015. Množičenje za slovar sodobnega slovenskega jezika. *Slovar sodobne slovenščine: problemi in rešitve.* Eds. Gorjanc, Vojko et al. Znanstvena založba Filozofske fakultete. Ljubljana.

Gantar, Polona. 2015. *Leksikografski opis slovenščine v digitalnem okolju*. Znanstvena založba Filozofske fakultete. Ljubljana. doi.org/10.4312/9789612377922.

Gantar, Polona; Kosem, Iztok; Krek, Simon. 2015. Discovering automated lexicography: the case of Slovene lexical database. *International journal of lexicography* 29/2. 200–225. doi.org/10.1093/ijl/ecw014.

Geyken, Alexander. 2019. The Centre for Digital Lexicography of the German Language: New Perspectives for Smart Lexicography. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography.* Eds. Kosem, Iztok; Zingano Kuhn, Tanara. Lexical Computing CZ s.r.o. Brno.

Gorjanc, Vojko; Gantar, Polona; Kosem, Iztok; Krek, Simon. 2017. *Dictionary of modern Slovene: problems and solutions*. Ljubljana University Press – Faculty of Arts. Ljubljana.

Kosem, Iztok; Husák, Miloš; McCcarthy, Diana. 2011. GDEX for Slovene. *Electronic Lexicography in the 21st Century: New applications for new users (Proceedings of eLex 2011).* Eds. Kosem, Iztok; Kosem, Karmen. Trojina, zavod za uporabno slovenistiko

– Institute for Applied Slovene Studies. Ljubljana. 151–159. www.trojina.si/elex2011/elex2011_proceedings.pdf (accessed 31 March 2020).

Kosem, Iztok; Krek, Simon; Gantar, Polona; Arhar Holdt, Špela; Čibej, Jaka; Laskowski, Cyprian Adam. 2018. Collocations dictionary of modern Slovene. *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts*. Eds. Krek, Simon; Čibej, Jaka; Gorjanc, Vojko; Kosem, Iztok. University Press – Faculty of Arts. Ljubljana. 989–997. doi.org/10.4312/9789610600961.

Krek, Simon; Laskowski, Cyprian Adam; Robnik Šikonja, Marko. 2017. From translation equivalents to synonyms: creation of a Slovene thesaurus using word co-occurrence network analysis. *Electronic lexicography in the 21st century: proceedings of eLex 2017 Conference*. Eds. Kosem, Iztok et al. Lexical Computing. Leiden. 93–109. https://elex.link/elex2017/proceedings/eLex_2017_Proceedings.pdf (accessed 31 March 2020).

Tavast, Arvi; Langemets, Margit; Kallas, Jelena; Koppel, Kristina. 2018. Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. *Proceedings of the 18th EURALEX International Congress: Lexicography in Global Contexts.* Eds. Krek, Simon; Čibej, Jaka; Gorjanc, Vojko; Kosem, Iztok. University Press – Faculty of Arts. Ljubljana. 749–761. doi.org/10.4312/9789610600961.

Rječnik suvremenoga slovenskog jezika: od slovenske leksičke baze do digitalne rječničke baze

Sažetak

Ideja sveobuhvatne rječničke baze koja uključuje sve razine jezičnoga opisa suvremenoga slovenskog jezika od morfološke i leksičke do sintaktičke prvotno je formulirana u okviru projekta *Sporazumijevanje na slovenskomu jeziku* (2008. – 2013.). U cilju ostvarenja ideje o stvaranju sveobuhvatne digitalne rječničke baze stvorene su dvije neovisne baze podataka: *Kolokacijski rječnik suvremenoga slovenskoga jezika* i automatski generiran *Tezaurus modernoga slovenskoga jezika*. Jedna od novina u obama rječnicima koncept je responzivnoga rječnika, koji uključuje masovnu podršku. *Digitalna rječnička baza* sadržava sve razine jezičnoga opisa: morfološku nadograđenu *Sloleksom*, izraznu s opisom konstrukcija višerječnih jedinica te sintaktičku s formalizacijom modela glagolskih valencija. Svaka od postojećih baza podataka sadržava specifične jezične podatke koji će biti uključeni u sveobuhvatnu *Slovensku digitalnu rječničku bazu podataka,* koja će sadržavati temeljni jezikoslovni opis slovenskoga jezika čiji korisnici mogu biti ljudi i strojevi.

Keywords: dictionary of modern Slovene, digital dictionary database, electronic lexicography, digital-born dictionary

Ključne riječi: rječnik suvremenoga slovenskog jezika, digitalna rječnička baza podataka, elektronička leksikografija, digitalni rječnik