The Dictionary of the Serbian Academy: from the Text to the Lexical Database

Ranka Stanković, Rada Stijović, Duško Vitas, Cvetana Krstev, Olga Sabo

The *Dictionary of the Serbian Academy*: from the Text to the Lexical Database

***Ranka Stanković*[1], *Rada Stijović*[2], *Duško Vitas*[1], *Cvetana Krstev*[1], *Olga Sabo*[2]**
[1]*University of Belgrade,* [2]*Institute for Serbian Language, Serbian Academy of Sciences and Arts*
E-mail: ranka.stankovic@rgf.bg.ac.rs, rada.stijovic@isj.sanu.ac.rs, vitas@matf.bg.ac.rs, cvetana@matf.bg.ac.rs, olga011@yahoo.com

Abstract

In this paper we discuss the project of digitization of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language*. Scanning and character recognition were a particular challenge, since various non-standard character set encoding was used in the course of the almost 60-year long production of the dictionary. The first aim of the project was to formalize the micro-structure of the dictionary articles in order to parse the digitized text of and transform it into structured data stored in relational lexical database. This approach is compatible with several standard structured forms and ontologies (TEI, LMF, Ontolex, LexInfo). A lexical database model was designed in compliance with these structured forms, following mostly the *lemon* model. Mapping of the lexical entry markers to LexInfo and TEI enabled export of the lexical data to the mentioned formats. A software solution for the dictionary text analysis, parsing and lexical database population was developed and tested on the first and the last published volumes of the dictionary (which contain 27,141 articles in total). An evaluation of the results shows that the developed model and software solution can be successfully used for the other volumes as well.

Keywords: computer lexicography, lexical database, language resources, dictionary, Serbian language

1 Introduction

The first volume of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language* (referred to as the *Dictionary of Serbian Academy* or *DSA*), prepared and compiled by the Institute for the Serbian Language of the Serbian Academy of Sciences and Arts, was published in 1959 in paper form. Out of 35 planned volumes, 19 volumes have been published with the 20th volume to be released soon. The material used for the dictionary covers written resources of the standard Serbo-Croatian language from the beginning of the 19th century to the present day, as well as about 300 word collections (provincial expressions, dialectical variations, etc.) of all Shtokavian dialects. The paper version of the dictionary has a complex microstructure that was designed at the time of the release of the first volume. Later, it was supplemented and described in a handbook.[1] The text itself is in basic Cyrillic alphabet, but in addition to this it contains accented vocals and various Church Slavonic, Latin and Greek characters.

It was first suggested that the production of the dictionary should be modernized many years ago (Sabo & Vitas 1989). However, only in recent years were these ideas revitalized, and various possibilities of updating the work on this vocabulary have since been considered (Vitas & Krstev, 2015; Ivanović et al. 2016). The digitization (which is also the topic of the present paper) of the published volumes and raw materials (lexicographic leaflets) began in 2016, and the first use of the two volumes that were

1 Упутство за обраду Речника, Београд: Институт за српск(охрватск)и језик САНУ (рукопис), 1959. и (допуњено) 2017 [*A Handbook for Dictionary Processing*, Belgrade: Institute for Serbo(-Croatian) language SASA (manuscript), 1959 and (supplement) 2017].

digitized and published was reported in Stijović et al. (2017), while the application for management of lexicographic leaflets is described in Stijović (2018). Out of 19 volumes, two were available as MS Word files, two as PDF files, and the others only in paper form. Unfortunately, neither MS Word nor PDF files could be used without further preprocessing, since non-Unicode character sets were used with non-standard accents and other forms of character encoding. All three available formats needed additional transformation: conversion, transliteration (where the Latin alphabet was used instead of Cyrillic) and extensive manual postediting. After the implementation of OCR, dictionary articles were corrected and formatted (bold, italic) manually to obtain the identical layout and formatting as in the printed version. The final correction of all 19 digitized volumes is nearly finished.

In this paper we will present the work that has been done so far in transforming the digitized text of the *Dictionary* into various standard structured formats and into a lexical database with the first aim to speed up the linear production process of the dictionary. This work makes it possible to use the lexical base of the dictionary for research purposes and for the production of various derived lexicographic products.

2 Related Work

Digital dictionaries ceased to be a novelty a long time ago. The majority of new dictionaries are produced (and in some cases exist only) in digital form. However, many significant lexicographic works that were produced in the past now need to be transformed into this format. Ever since some initial retro-digitization projects, such as the transformation of the *Oxford English Dictionary* (Berg et al., 1988), the transformation from an unstructured to a structured text was recognized as the main task of such endeavors. To do this, the text of a dictionary has to be parsed and the structure of the articles has to be formalized. For the representation of this formal structure markup languages are used, preferably standard ones that support interchange and merging (Lemnitzer et al., 2009). At this point, retro-digitized and digital-born dictionaries meet, since both types should preferably use the same or compatible formal structure and markup language.[2] This development led to further linking of lexical data and their integration with semantic resources, such as ontologies (McCrae et al., 2011).

The *DSA* is rather special compared to similar dictionaries for other languages: its significant part has already been compiled and published, but still needs to be digitized; on the other hand, the dictionary is not finished, and a lot of work remains to be done. These two tasks need to be synchronized in order to obtain a homogenous work as the final product. Moreover, although the digitization of the *Dictionary* is currently lagging, we would like to catch up on the lost time by using up-to-date technology.

3 Model of Lexical Database

3.1 Formalization of the structure of dictionary articles

The first phase of the conversion of the *DSA* from the text form (unstructured text) into the lexical base (structured text) consisted of a thorough analysis of formatting conventions that were used for typesetting dictionary entries as well as of the identification of triggers (such as special words, abbreviations or punctuation marks) used to introduce specific information. Conventions and triggers were used to identify basic information presented in the dictionary. This analysis enabled us to recognize the following entry structure:

2 For instance, Ahačić (2015) presents a Slovenian Dictionary Portal that collects information from 22 dictionaries, dating from the 16th century to the present day. Some of these dictionaries were transformed into XML format, while other were developed in it.

1) headword group;
    1. headword lemma;
    2. grammatical data;
    3. related words (lexical entries);
2) grammatical data;
3) etymology – an element from the closed set of abbreviations for the names of languages is used to introduce the information on etymology, for instance *грч*. for *грчки* 'Greek' or *фр*. for *француски* 'French';
4) sense – the individual meanings of a headword are marked with Arabic numerals or, if the meanings are close, with lowercase letters. If the headword is a verb, its non-reflexive and reflexive forms are marked by Roman numerals I and II.
    1. terminological markers – predefined terminological abbreviations are used to indicate the scientific domain in which a particular sense of a lemma is used (for instance, *геол*. for *геологија* 'geology' or *фил*. for *филозофија* 'philosophy');
    2. the linguistic and stylistic tags – qualifiers of predefined linguistic and stylistic values (for instance, *покр*. for *покрајински* 'provincial', *арх*. for *архаичан* 'archaic', *пеј*. for *пежоративан* 'pejorative');
    3. related words (lexical entries) – references to other lexical entries (for instance, preferred forms) are introduced after appropriate abbreviations: *исп*. for *испоредu* 'compare';
    4. definitions are descriptive or referential (*в.* for *види* 'see'), in rare cases synonyms;
    5. definitions are supplemented by the lists of:
        1. synonyms (after abbreviation *син*. for *синоним* 'synonym');
        2. antonyms (after abbreviation *супр*. for *супротан* 'antonym');
        3. related words;
    6. examples:
        1. the text of the example
        2. the bibliographic reference (in parenthesis);
5) multiword expressions (syntagmatic and phraseological – they are listed in the separate paragraph beginning with the abbreviation *Изр*. for *Израз* 'phrase'; and
6) proverbs – they are also listed in the separate paragraph, beginning with the abbreviation *НПосл*. for *Народна пословица* 'vernacular proverb'.

Some elements of this structure may appear at different positions and levels, such as 'related words'. Grammatical data can contain various information, depending on the lemma's part-of-speech: some may refer to the lemma, others to the headword. Both high- and low-level elements are optional and repeatable, except for the headword group itself. The typographic conventions and triggers as applied to nouns are summarized in a simplified way in Table 1, while the graphical outline of the basic structure of an article in the *Dictionary* is presented in Figure 1. The same entry has been taken as an example in Table 1 and Figure 2, which illustrates the result of the parsing process.

3.2 Dictionary markers

Beside various semantic, accentual and grammatical (phonetic, morphological and, more recently, syntactic) information, the *DSA* also includes indications of the normative, functional, stylistic and socio-historical status of the lexical entries, as well as their spatial and temporal scope and domain of use. Apart from other lexicographical and technical procedures, various markers are used to denote the status of the lexemes and to detail the rules of their use. They are placed at different positions in the articles of the *Dictionary* following strict rules. A total of 371 such markers, in abbreviated forms, is used, and they are all listed at the beginning of each printed volume. For the purpose of the

digitization, a meticulous systematization of all such markers, in terms of the information they convey and their position in the articles of the *Dictionary*, was performed.

A feature structure, as a general-purpose data structure which identifies and groups together individual features, each of which associates a name with one or more values, was mapped with the help of the aforementioned abbreviations used in the dictionary. The existing 371 abbreviations (markers) were mapped as data category values with 30 data categories, and further grouped in data-category sets. Feature structures represent the interrelations among various pieces of information and provide a metalanguage for the generic representation of the analyses and interpretations.

Table 1: The typographic conventions and triggers as applied to nouns.

Element		Example	Typography	Trigger begin	Trigger end
Headword group					
	lemma	па̀леоцēн	<nl> bold		comma or trigger begin
	gramm. data	-а		hyphen	
	lemma	палеòцēн	<nl> bold	и	comma or trigger begin
	gramm. data	-ена		hyphen	
gramm. data		м		item in a list	
Etymology		palaiós kainós		Open parenthesis + item in a list, e.g. „(грч.“	closing parenthesis
Sense				1, 2, 3 or а, б, в or I, II or trigger begin	
	terminological markers	геол.		item in a list	trigger begin
	linguistic/markers	/		item in a list	trigger begin
	related words	/		Some punctuation marks; item in a list	trigger begin
	definition	прва, најстарија епоха палеогена.	italic		trigger begin
	synonyms, antonyms, related	/		item in a list	
Example	example text	Формације геолошке се даље дијеле...		dash	
	Bibliographic references	Д-П1, 17		Open parentheses	Closing parenthesis
MWE		/		*Изр.*	
Proverbs.		/		НПосл.	

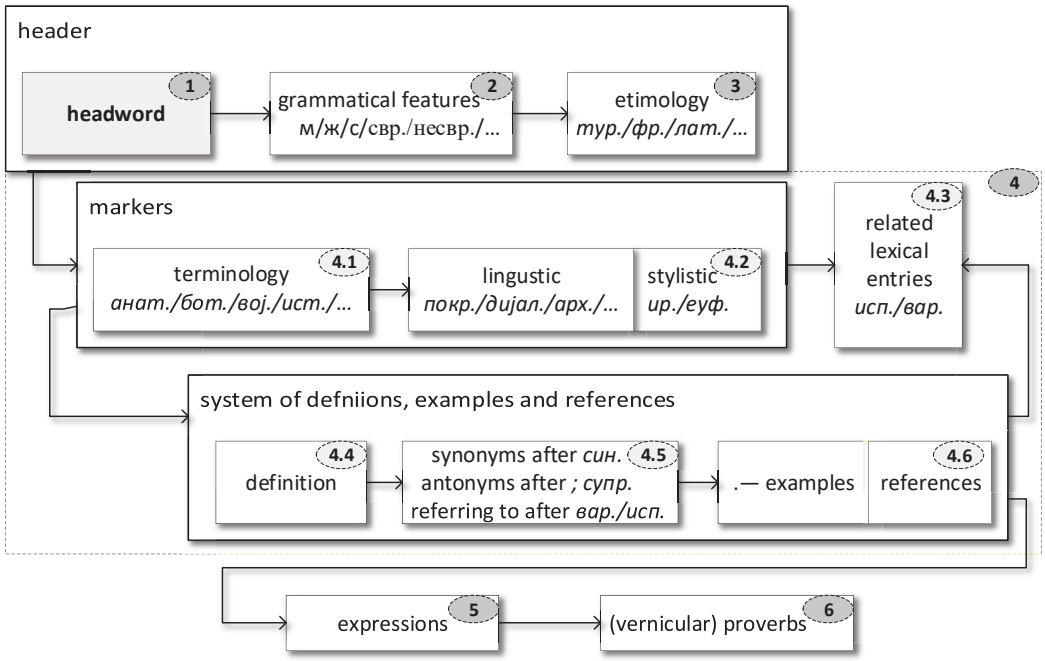

Figure 1: The microstructure of dictionary articles.

4 The transformation from the dictionary article text form to the lexical database

The guidelines for dictionary writing were used to define the rules for the segmentation of the dictionary articles, the pattern recognition, and the alignment of the recognized markers with the predefined categories, as described in the previous section. The dictionary article units that were recognized were marked with XML tags, in accordance with the predefined scheme and imported into the relational database. The model of the lexical database was inspired by LMF (Calzolari et al. 2013), TEI[3] and *lemon* (McCrae et al. 2011). The current trends seem to be consistent with the idea of an ecosystem, where different standards can coexist and mutually enrich each other (). We have experimented with partial transformations of our XML documents to these models in order to find the most suitable solution.

MS Word documents formatted identically as the print versions were input to the automatic segmentation procedure. One thus prepared dictionary article is presented on the left side of Figure 2, while the result of the segmentation of the same dictionary article is presented on the right. One of the additional functions of the developed software is the consistency check that reports any occurrence that does not comply with the syntactic rules for the predefined article structure. This information is then used for the manual postediting of the digitized text.

Within this research, the partial alignment of the XML tag set, defined for the DSA with the TEI dictionary module, was made. For instance, the <*gen*> element is used to denote the grammatical gender, while <*usg type = "dom"*> refers to the terminological field, and <*def*> to the definition. Citations are tagged with <*cit*>, and bibliographic references with <*bibl*>; in the content of these elements some additional phrases are tagged, such as the names of locations (using the <*placeName*> tag) or authors (using the <*author*> tag). Similarly, comparison and partial alignment of the DSA tag

3 http://www.tei-c.org/

set was done with Ontolex[4] and LexInfo[5], but a more precise and detailed alignment is envisaged. The dictionary article from Figure 2 is represented in the TEI compliant form in Figure 3.

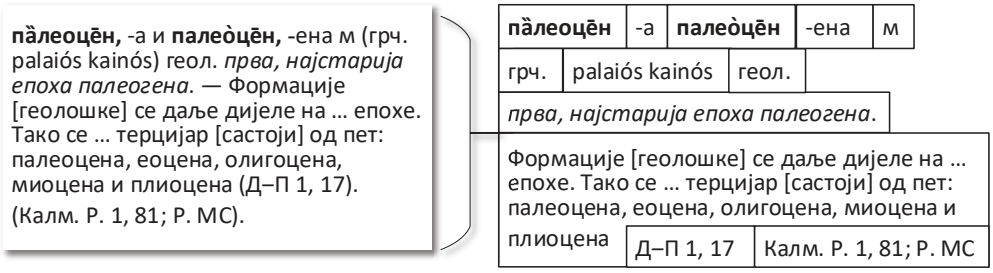

Figure 2: An example of the automatic dictionary article segmentation.

The information that was registered in the dictionary explicitly, such as the domain, the stylistic use, the etymology, and so on, was mapped with the help of the data categories and their values; however, some grammatical information was not explicitly encoded. For example, part-of-speech (POS) is rarely encoded explicitly, but rather through an indirect indicator. For instance, the gender mark (*м* masculine, *ж* feminine, *с* neuter) indicates the grammatical category of nouns, the aspect type mark (*свр*. for *свршен* 'perfective', *несвр*. for *несвршен* 'imperfective') is the indicator of verbs, while adjectives are given in all three grammatical genders in the nominative singular (e.g. *активан, -вна, -вно* 'active'). The set of rules was produced that calculates the POS where it is not explicitly mentioned, and their application yielded correct POS information for more than 95% of all lexical entries.

```
<entry n=”3971”>
   <form type=”lemma”><orth> па̀леоцēн </orth>,</form>
   <form type=”inflected”>-a</form>
   <form type=”lemma”><orth> палеò̯цēн </orth>,</form>
   <form type=”inflected”>-a</form>
   <gramGrp><gen>м</gen></gramGrp>
   <sense><etym>(<lang>грч.</lang> palaiós kainós) </etym><usg type=”dom”> геол.</usg>
      <def> прва, најстарија епоха палеогена.</def>
      <cit> Формације [геолошке] се даље дијеле на … епохе. Тако се … терцијар [састоји] од пет: палеоцена,
еоцена, олигоцена, миоцена и плиоцена <bibl>( Д–П 1, 17).</bibl>( Калм. Р. 1, 81; Р. МС) </cit>
   </sense>
</entry>.
```

Figure 3: The example of a dictionary article using TEI compliant tagging.

Our main goal is to produce a central lexical database that will enable multiuser management of the lexical data and provide access to the content of the volumes that were already published. For the development of the lexical database model for the *DSA*, a similar approach was used as in Stanković et al. (2018) for the Serbian morphological electronic dictionary. The main class, in the core of this dictionary model, is *LexicalEntry*, representing a headword of the dictionary article, which encompasses the set of senses that are associated with this headword. The *LexicalRelation* class relates lexical variants (for instance, *па̀леоцēн* and *палеò̯цēн* 'paleocen' in our example), full forms and their abbreviations (*др* for *доктор* 'doctor'), orthographic variants and different pronunciations (Ekavian *sneg* and Ijekavian *snijeg* 'snow'). *LexicalSense* is used to represent a particular sense of a lexical entry, and to

4 https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

5 http://www.lexinfo.net/ontology/2.0/lexinfo

link a lexical entry with a set of senses. Each sense can be related to its own set of markers (outlined in Section 3.2) through the *SenseProperties*. These markers are controlled by the internal thesaurus of data categories, as outlined in Stijović and Stanković (2017).

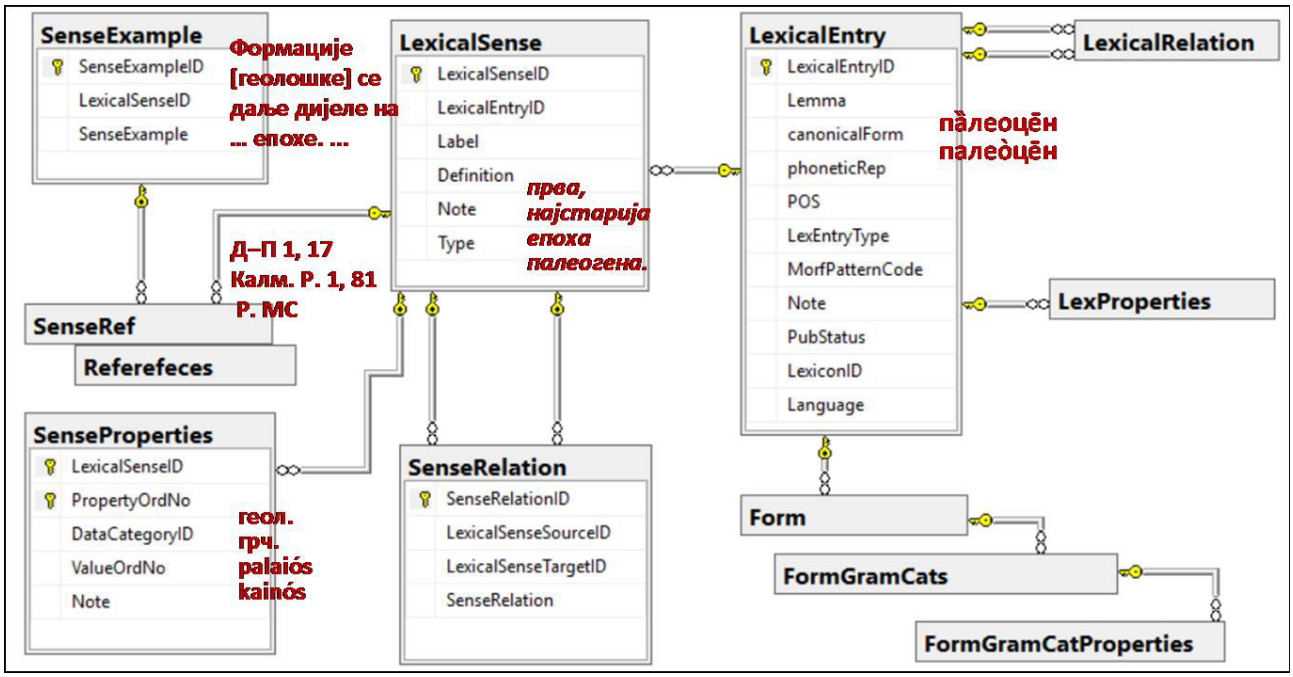

Figure 4: Database model with mapped dictionary article segments.

For languages with complex morphology, such as Serbian, information about inflection is very important. In our model we used the class *Forms* to store the information related to the inflected forms, while the grammatical categories are assigned to the *LexProperties*. At this moment, only information found in the *DSA* is stored in the table *Forms* – sometimes this information represents a selection of inflected forms, sometimes only inflectional endings. The *SenseRelation* is used to connect various senses of lexical entries, while the *SenseRef* and *SenseExample* contain information about provenance and use. The class *References* contains metadata about the bibliographic information from the dictionary corpus. The set of markers is partially aligned with the TEI elements (and attributes) and LexInfo in order to relate the lexical data to other resources and provide automatic production of the dictionary in different forms and formats. Figure 3 illustrates a part of the database model with a few examples of structured data.

5 Results and discussion

The procedure that was presented in this paper was applied to the digitization of the 1st (1959) and 19th (2014) volumes of the *DSA*. As a result, the structured documents were obtained in standardized formats and they were subsequently loaded into the lexical database. The automatic procedure recognized, structured, annotated and stored in the lexical database 15,988 dictionary articles from the 1st volume and 11,153 from the 19th.

Two processed volumes were compared regarding the POS of headwords, and the absolute and relative frequencies were compared: 10,633 lexical entries were recognized as nouns in the 1st volume (66.2% of the total number of entries) and 7,808 (69.7%) in the 19th, 1,364 (8.5%) entries were recognized as verbs in the 1st volume and 1,192 (10.6%) in the 19th, while 2,654 (16,5%) entries were

recognized as adjectives in the 1st volume and 1,391 (12,4%) in the 19th volume. A small number of entries was not labelled with POS: 607 (3.8%) and 521 (4.7%) in 1st and 19th volumes, respectively, due to the lack of information in the *Dictionary* itself or the lack of appropriate rules; the improvement of the set of rules for POS detection is underway (Stijović & Stanković, 2017). The 1st volume has more dictionary articles than the 19th (15,988 vs. 11,153), but at the same time less tokens (1,722,483 vs. 1,987,504) and formal words (551,030 vs. 672,004). The 1st volume refers to 2,105 different sources with 7,127 examples, while the 19th refers to 6,037 different sources with 28,725 examples.[6] Some other comparative differences between 1st and 19th volume are presented in Figure 5.

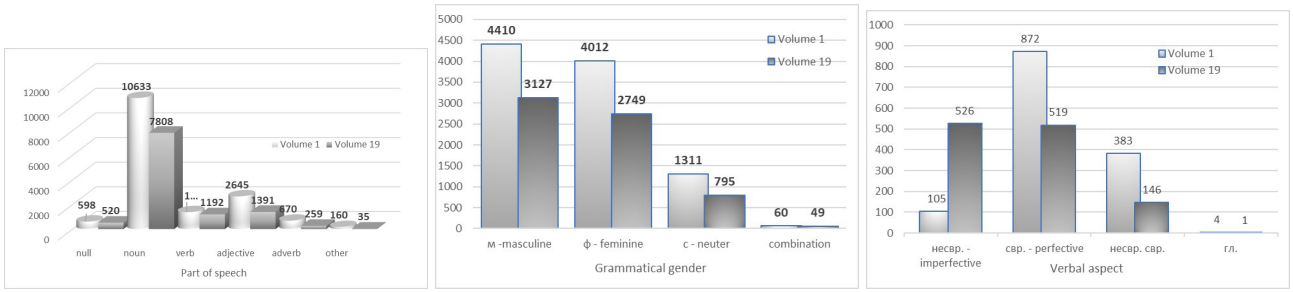

Figure 5: The comparison of two volumes: by part-of-Speech (left), nouns by grammatical gender (middle), and verbs by aspect (right) by lexical entries types from data in lexical database.

6 Conclusion

In this paper we discussed the challenges and results of the digitization of the *DSA*. The results presented in this paper are restricted to the processing of the 1st and 19th volumes, but the digitization of all previously published volumes is in progress. So far, the *Dictionary* was produced linearly. In the future, however, the linear processing of entries may be abandoned, as this could accelerate the production process. The supplements to the already published volumes could also be produced. Once the *DSA* is fully populated, users with different levels of accessibility will be able to search through its lexical database. It is also envisaged for the basic data will be open to the general public.

References

Ahačič, K., Ledinek, N., & Perdih, A. (2015). Fran: The Next Generation Slovenian Dictionary Portal. In *Natural Language Processing, Corpus Linguistics, Lexicography. Eight International Conference Bratislava, Slovakia,* pp. 21-22.

Berg, D. L., Gonnet, G. H., & Tompa, F. W. (1988). *The New Oxford English Dictionary Project at the University of Waterloo* (pp. 2-7). UW Centre for the New Oxford English Dictionary.

Calzolari, N., Monachini, M., Soria, C. (2013). LMF – Historical Context and Perspectives, in: LMF Lexical Markup Framework, Eds: G. Francopoulo, P. Paroubek, John Wiley & Sons, Inc.

Ivanović, N., Jakić, M., Ristić, S. (2016). Građa Rečnika SANU – potrebe i mogućnosti digitalizacije u svetlu savremenih pristupa, u: S. Ristić i dr. (ed.), Leksikologija i leksikografija u svetlu savremenih pristupa, Beograd: Institut za srpski jezik SANU, pp. 133–154. [The material of the Dictionary of the SANU - the needs and possibilities of digitization in the light of contemporary approaches (in Cyrillic)].

Lemnitzer, L., Romary, L., & Witt, A. (2009). Representing human and machine dictionaries in Markup languages. In *arXiv preprint arXiv:0912.2881*.

6 The 19th volume contains fewer articles but they have a more complex semantic structure, which accounts for more meanings, and consequently more examples.

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with *lemon*. In *Extended Semantic Web Conference* Springer, Berlin, Heidelberg, pp. 245-259.

Monachini, M. & Khan, A. F. (2018). Towards the Construction of a Lexical Data and Technology Ecosystem: The Experience of ILC-CNR. In Proceedings of the LREC 2018 Workshop "Globalex 2018 – Lexicography & WordNets", Eds: Ilan Kernerman, Simon Krek, Miyazaki, Japan, pp. 52-54.

Sabo, O., Vitas, D. (1998). Mogućnost osavremenjivanja izrade rečnika na primeru Rečnika srpskohrvatskog književnog i narodnog jezika SANU i Instituta za srpskohrvatski jezik. In *IV međunarodni naučni skup „Računarska obrada jezičkih podataka",* Portorož: Institut Jožef Stefan, pp. 375–384. [Possibility for modernizing the development of the dictionary on the example of the Dictionary of the Serbo-Croatian literary and vernacular language SASA and the Institute for Serbo-Croatian]

Stanković, R., Krstev, C., Lazić, B., Škorić, M. (2018) Electronic Dictionaries – from File System to *lemon* Based Lexical Database, In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science* (in print)

Stijović, R., Sabo, O., Stanković, R. (2017). Rečnik SANU kao baza terminoloških rečnika (na primeru Rečnika kulinarstva) u: Piper, P., Jovanović, V. (eds.), Slovenska terminologija danas, Beograd: Srpska akademija nauka i umetnosti, 2017, pp. 229–241. [The dictionary of the SASA as a basis of terminology dictionaries (on the example of the Culinary Dictionary) (in Cyrillic)]

Stijović, R. (2018). Građa Rečnika SANU – blago koje treba sačuvati (o digitalizaciji listića), In *Naš jezik XLVI-II/3–4, Beograd: Institut za srpski jezik SANU*, pp. 201–207. [The structure of the Dictionary of the SANU - the goods to be preserved (on the digitization of the leaflets) (in Cyrillic)]

Stijović, R., Stanković, R. (2017) Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU, In *47. Naučni sastanak slavista u Vukove dane, Beograd* (Međunarodni slavistički centar. Filološki fakultet) (in print). [Digital edition of the SASA Dictionary: a formal description of the microstructure of the SASA Dictionary (in Cyrillic)]

Vitas D., Krstev C. (2015) Nacrt za informatizovani rečnik srpskog jezika, In *Naučni sastanak slavista u Vukove dane - Srpski jezik i njegovi resursi: teorija, opis i primene, Vol. 44/3*, Međunarodni slavistički centar, Beograd, pp. 105-116. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic)]

Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003, #178003 and #178009.