The Dictionary of Contemporary Serbian Language (RSSJ): Advanced Automation and Other Challenges

Ranka Stanković[1], Rada Stijović[1], Mihailo Škorić[1], Cvetana Krstev[1]

[1]JERTEH - Language Resources and Technologies Society, Djusina 7, Belgrade, Serbia
E-mail: ranka@jerteh.rs, stijovicr@gmail.com, mihailo@jerteh.rs, cvetana@jerteh.rs

Abstract

This paper introduces the Dictionary of Contemporary Serbian Language (RSSJ), an ongoing large-scale digital lexicographic project designed to serve both human users via web and mobile applications and machines through APIs. Coordinated by the diaspora association "Gathered around the Language" and the Society for Language Resources and Technologies (JeRTeh), RSSJ aims to produce a dictionary of approximately 50,000 frequently used words, reflecting vocabulary used over the past fifty years across diverse functional styles. The headword list is automatically extracted from corpora (SrpKor2013, SrpKor2021), then manually curated and enriched with data from the LeXimirka database. The project implements advanced automation at multiple stages, employing language models and static embeddings (Word2Vec, FastText, Dict2Vec) to identify synonyms, while large language models assisted in generating draft definitions. Additional methods include automated extraction of collocations, syntactic patterns, and exemplary usage via GDEX algorithms, all managed within a DMLex-inspired PostgreSQL data model. The custom web interface enables seamless integration of dictionary editing and corpus querying. Preliminary results demonstrate that automated drafting accelerates to some extent dictionary development, requiring at the same time lexicographers to adopt more dynamic, data-driven workflows and redefine traditional lexicographic practices.

Keywords: dictionary, Serbian language, lexicography, lexicographic database, natural language processing, large language models, word embeddings

1. Introduction

This paper outlines the motivation, concept, implementation challenges, and solutions for the Dictionary of Contemporary Serbian Language (RSSJ) project. Initiated by the "Gathered around the Language" diaspora association and supported by the Society for Language Resources and Technologies (JeRTeh),[1] the project aims to produce both a comprehensive dictionary as a digital lexicographic database for human use via web and mobile applications and machine use via APIs. The goal is to create a dictionary of approximately 50,000 entries, adhering to contemporary lexicographic and IT standards.

Despite significant advances in digital lexicography globally, there remains a notable lack of modern, comprehensive online dictionaries for the Serbian language. Existing resources contain quite a few outdated headwords or specific senses that are no longer used in contemporary language, limited in scope, or lack features such as real-time updates, API access, and advanced search functionalities needed by both language learners and researchers. This gap highlights the urgent need for an up-to-date, user-friendly, and

[1] https://jerteh.rs/

technologically advanced Serbian dictionary, serving as the key motivation and inspiration for the RSSJ project. By addressing this deficiency, the project aims to bring Serbian lexicography in line with current international standards and digital best practices.

We highlight the synergy between human expertise, language resources and AI-driven automation in modern lexicography. Section 2 brings the overview of automated dictionary development approaches as the background of research. The methodology of the approach is given in Section 4 and the developed application for dictionary writing in Section 4. The paper will conclude by analyzing the benefits and drawbacks of this AI-assisted approach, as well as discussing the integrated development environment and its usage modalities (Section 6). The limitation of the approach and plan for improvements will be given in Section 6. Preliminary results indicate that using automatically drafted entries speeds up development to some extent. However, this shift in the dictionary development process shows that trained lexicographers are no less needed than before. At the same time, they have to stay open-minded and move away from the traditional linear workflow and embrace a more flexible, content-driven approach.

2. Automated Dictionary Development Approaches Overview

In recent years, dictionary compilation has undergone a paradigm shift due to advances in artificial intelligence (AI), machine learning, and large-scale language resources. Traditionally, dictionary making was a highly manual and time-consuming process (Atkins & Rundell, 2008). However, contemporary projects increasingly leverage computational methods to automate tasks such as headword extraction, sense induction, and example selection, fundamentally transforming lexicographic practice (Kosem et al., 2020; Tiberius et al., 2024; Klosa-Kückelhaus & Tiberius, 2024; de Schryver, 2024).

The integration of large corpora into dictionary development allows lexicographers to have direct access to usage examples, leading to more representative and up-to-date lexicons. Tools such as Sketch Engine (Kilgarriff et al., 2004, 2014) and NoSketchEngine[2] facilitate automated extraction of frequency data, collocations, and usage examples, supporting lexicographers in creating data-driven entries.

AI and machine learning models are now routinely used to identify headwords, cluster senses, extract candidate definitions, and suggest synonyms and antonyms (Navigli & Velardi, 2010). Static embeddings (e.g., Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017)) and, more recently, contextual embeddings (e.g., BERT (Devlin et al., 2019) have enhanced automatic synonym and semantic relation discovery, as well as the identification of new word senses (Pilehvar & Camacho-Collados, 2021)).

Recent studies demonstrate that LLMs such as GPT-3, ChatGPT, and Gemini can generate draft definitions and usage examples, and even provide stylistic adaptation for different dictionary types (Krek, 2024).

Although AI-generated content accelerates lexicographic workflows and reduces costs, human validation and editing remains crucial to ensure linguistic quality and cultural appropriateness.

[2] https://www.sketchengine.eu/nosketch-engine/

Algorithms like Good Dictionary Example (GDEX) (Kilgarriff et al., 2008) have become standard for automatically selecting exemplary sentences from corpora, further reducing the manual effort required by lexicographers.

Modern dictionary projects now often employ integrated platforms that combine corpus access, automated annotation, and lexicographic editing (Kosem et al., 2021). Data models such as DMLex[3] (Měchura, 2024) enable structured and interoperable dictionary databases suitable for human and machine use.

Despite these advances, challenges remain, particularly regarding the balance between automation and editorial oversight, the handling of low-resource languages, and the evaluation of AI-generated content for accuracy and cultural fit (Gantar, 2024).

3. Methodology

Vitas & Krstev (2012) presented an overview of resources specifically constructed for the processing of Serbian, namely, language corpora, morphological electronic dictionaries and semantic networks. They presented the advantages of processing Serbian corpora using electronic dictionaries. On the basis of these results, they proposed the blueprint for the computerized dictionary of the Serbian language (Vitas & Krstev, 2015) that relies on the development of a computerized infrastructure for the study and processing of the Serbian language, to serve as a fundamental resource for future Serbian lexicography. The aim was to demonstrate how corpora can be successfully exploited using high-recall tagging, which differs significantly from the mainstream approach based on one-to-one tagging prior to any processing.

A significant step towards the achievement of these goals was the development of *LeXimirka*, a sophisticated lexicographic database and web application designed for developing, managing, and exploring lexical data, originally created for the above mentioned Serbian language resources (Krstev, 2008; Stanković et al., 2018a; Lazić & Škorić, 2019). It manages morphological and semantic information, rule-based automatic linking of lexical entries (e.g., variant forms, different pronunciation, derivational entries, multilingual mappings) providing multi-user collaboration with the control of entry editing, corpus-driven automatic vocabulary enrichment, and linking entries across languages. LeXimirka data model is inspired by standards Ontolex-Lemon[4] and LMF[5] and integrates closely with corpora and NLP pipelines, enabling efficient, data-driven lexicographic workflow. Its relational model and web interface make it well-suited for both human lexicographers and downstream NLP applications.

Serbian WordNet (SWN)[6] (Krstev et al., 2004; Stanković et al., 2018b) was initially developed as a part of the BalkaNet project (Tufis et al., 2004), which covered a number of Balkan languages, as well as within the broader EuroWordNet initiative for European languages. Alignment across languages has been achieved through the Inter-Lingual Index

[3] https://www.oasis-open.org/2025/05/29/dmlex-approved-as-oasis-standard/, https://docs.oasis-open.org/lexidma/dmlex/v1.0/OS/dmlex-v1.0-os.html

[4] https://www.w3.org/2016/05/ontolex/

[5] ISO standard, Language resource management – Lexical markup framework https://www.iso.org/standard/37327.html, deprecated 2019, the new version is now available, but this one was valid at the time LeXimirka database was established.

[6] https://wn.jerteh.rs/

(ILI), which was established to connect semantically similar concepts in different languages based on the Princeton WordNet (PWN).

Figure 1 presents the workflow for the development of RSSJ dictionary, starting with the initial steps: conceptualization, planning, selection of resources, microstructure definition, and followed by the most important steps that govern the development: automatic information extraction from existing resources and data generation using prompt engineering.

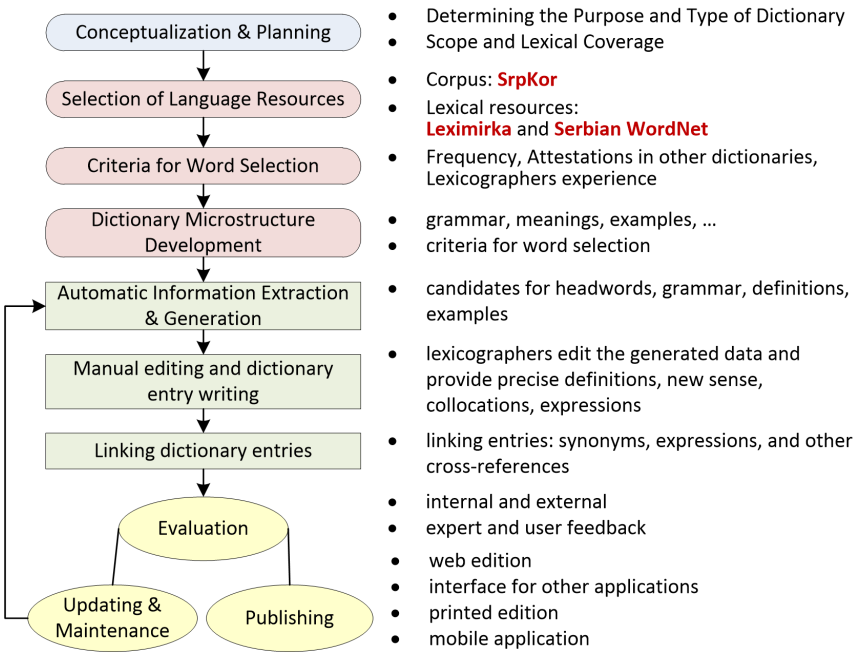

Figure 1: The RSSJ workflow

The microstructure of the RSSJ dictionary (Figure 2) refers to the internal organization of individual entries, providing comprehensive and systematic lexical information for each headword. Each entry typically includes the canonical form of the word (lemma), with Ekavian and Ijekavian variants when applicable, and occasionally other types of variants. The part of speech is clearly indicated, with additional grammatical subcategorization. Selected inflectional forms or suffixes are provided, although the full paradigm is not listed. Definitions are presented concisely in accordance with contemporary lexicographic standards, ensuring that a broad audience can understand them. Short usage examples, usually drawn from large corpora, illustrate the word in contemporary and standard contexts. Entries may also include collocations and typical syntactic patterns to highlight standard usage, as well as lists of (near) synonyms, or related expressions where appropriate. Additional information on domain, register, or temporal status may be included to enhance understanding. Cross-references direct users to related entries or multi-word expressions.

In parallel with the definition of the microstructure, the criteria for word selection were established. It was decided to base the selection process on frequencies derived from corpora as the evidence of the words' current use, attestations in e-dictionaries/Leximirka, printed Serbian dictionaries, inclusion in Serbian WordNet, and the judgment of lexicographers. The RSSJ dictionary's headword candidates are extracted automatically from the large corpora of contemporary Serbian, SrpKor2013 and SrpKor2021 (Vitas et al., 2025) and manually

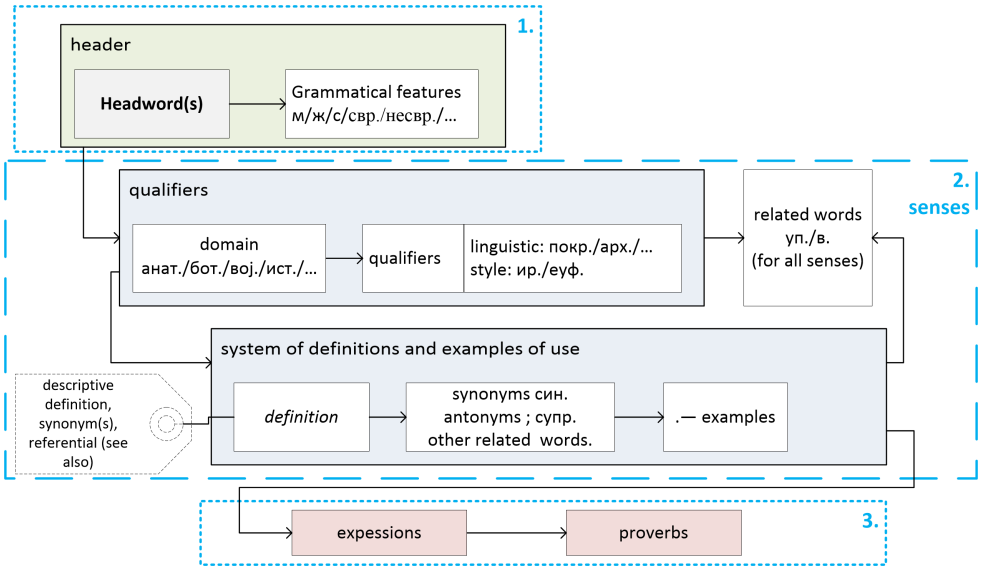

Figure 2: The RSSJ microstructure

evaluated by lexicographers. It encompasses the vocabulary across various styles (literary, scientific, journalistic, administrative, conversational) used approximately over the last fifty years. A fifty-year time frame is generally sufficient to provide a representative snapshot of contemporary language usage while minimizing the inclusion of outdated vocabulary and obsolete meanings that have fallen out of use. However, the most prominent authors from earlier periods, such as Ivo Andrić, have also been included, as their language is both rich and valuable, and their works remain widely read and influential in contemporary Serbian culture. Figure 3 presents a list of headword candidates (a canonical form and its part of speech (PoS)), accompanied by their corpora frequencies (SrpKor2013, SrpKor2021 and average), and their status in five selected dictionaries – indicators that show whether they are listed as headwords in them – all of which should help lexicographers decide which candidates to include in the dictionary.

Row	Rang	cirLemma	Da	POS	Sub	POScir	Komentar	Homogr	mall	FSK	FSK	FSK	Fre	Srp	RM	RR	WN	Dic
31221	1000	његов	☑	PRO	☐	зам.		null	3	166727	113864	129805	136799	1	1	0	0	0
31222	53000	његовање	☐	N	☑	им.		null	1	0	12	0	4	1	0	0	0	0
31223	41000	његовати	☐	V	☑	гл.		null	4	0	56	0	19	1	1	1	0	0
31224	53000	његошевски	☑	A	☐	прид.		null	3	0	12	0	4	1	1	0	0	0
31225	38000	његушки	☑	A	☐	прид.		null	1	0	24	55	26	1	0	0	0	0
31226	50000	њедра	☐	N	☑	им.		null	4	0	19	0	6	1	1	0	0	0
31227	40000	њежан	☐	A	☑	прид.		null	1	0	61	0	20	1	0	0	0	0
31228	51000	њежност	☐	N	☑	им.		null	1	0	16	0	5	1	0	0	0	0
31229	12000	њезин	☑	PRO	☐	зам.		null	3	456	582	411	483	1	1	0	0	0
31230	41000	Њемац	☐	N	☑	им.		null	4	1	28	27	19	1	1	0	0	0
31231	17000	њемачки	☐	A	☑	прид.		null	1	3	713	0	239	1	0	0	0	0
31232	1000	њен	☑	PRO	☐	зам.		null	4	78268	53162	87158	72863	1	1	0	0	0
31233	5000	њива	☑	N	☐	им.		null	3	1716	1422	3698	2279	1	1	0	0	0
31234	30000	њивица	☑	N	☐	им.		null	3	41	20	110	57	1	1	0	0	0
31235	53000	њивски	☑	A	☐	прид.		null	1	0	12	0	4	1	0	0	0	0
31236	23000	њин	☑	PRO	☐	зам.		null	3	280	74	0	118	1	1	0	0	0
31237	53000	њиов	☐	PRO	☐	зам.		null	1	0	11	0	4	1	0	0	0	0
31238	55000	њисак	☑	N	☐	им.		null	4	5	0	0	2	1	1	0	0	0
31239	55000	њискати	☑	V	☐	гл.		null	4	5	0	0	2	1	1	0	1	0
31240	51000	њихало	☐	N	☐	им.		null	1	15	0	0	5	1	0	0	0	0
31241	34000	њихање	☐	N	☐	им.		null	2	39	20	55	38	1	0	0	0	0
31242	17000	њихати	☑	V	☐	гл.		null	6	321	108	247	225	1	1	0	1	0
31243	1000	њихов	☑	PRO	☐	зам.		null	3	89991	80925	126902	99273	1	1	0	0	0
31244	51000	њиштати	☑	V	☐	гл.		null	4	15	0	0	5	1	1	0	0	1

Figure 3: The headword candidate list

The selected set of candidates was supplemented by additional grammatical information from the LeXimirka lexical database (Krstev, 2008; Stanković et al., 2018a; Lazić & Škorić, 2019). The RSSJ database do not include full inflectional paradigm, but rather an explanation that is similar to existing Serbian (paper) dictionaries. For example, for the verb акцентовати 'to accent' the grammatical information is:

> àкцентовати grammar is: -тујēм свр. и несвр.

The blueprint of 58K dictionary entries was automatically compiled and imported into lexical database implemented in PostgreSQL. During manual review of the offered list some headwords were excluded, some new were introduced, while some doublets were merged.

A cornerstone of the RSSJ (Stijović et al., 2025) development is a heavily data-driven approach, leveraging diverse computational methods. Beyond the SrpKor corpora and LeXimirka, this includes language models developed by JeRTeh and resources from the TESLA (Text Embeddings – Serbian Language Applications) project. Following the approach for aligning LLM-generated free association norms with human data (Abramski et al., 2025), we employ a similar comparative framework for synonym generation. Specifically, static embeddings like Word2Vec, FastText, and Dict2Vec from TE-SLA GitHub[7] (Stanković et al., 2025) are used to automatically identify candidates for synonyms. Since this is not a thesaurus and focus is not in listing all synonyms and related words, prepared lists are not imported into the database, but offered to lexicographers as a supplementary resource. Table 1 gives some examples from the dataset of synonyms: synonyms are presented for a dictionary headword that were retrieved from various lexical resources (the third column) and synonyms generated by GPT-4.1 (the last column). It can be seen that some of the generated candidates are listed in dictionaries as well (underlined), some are potentially useful, while some of the examples offered would be difficult to accept (candidates for the adjective 'акцијски' (actional)); moreover, there are also some non-existing words 'овогтренутка' and 'овогчаса' (*thismoment*) obtained by concatenation.

The project actively experiments with cutting-edge AI for lexicographic tasks: large language models such as Gemini 2.0 and GPT 4.0 in the first run, and later Gemini 2.5 and GPT 4.1 are utilized to generate draft definitions, trying to speed up to some extent the initial writing process. These efforts relied on results from automatic extraction methods and prompt engineering with Gemini 2.0 and GPT-4.0 models, which were among the most advanced available to us during the period of experimentation (December 2024 – February 2025). By leveraging existing Serbian dictionaries, various approaches for modeling and extracting definitions were systematically analyzed. The study in (Stijović et al., 2021) demonstrated that a substantial proportion of definition structures can be systematically modeled by grouping definitions according to type (in this case PoS) and utilizing template-based approaches. Furthermore, the automation process included experiments with corpus analysis and prompt engineering techniques to identify potential new meanings for existing headwords.

Entries are processed using contemporary lexicographic techniques, with prompts tailored to ensure that definitions are easily understandable to a wide audience. We grouped words by part of speech and then prepared prompts based on definition templates specific to

[7] https://huggingface.co/te-sla, https://huggingface.co/jerteh

Headword	**PoS**	**Synonyms (lexical resources)**	**Synonyms (GPT-4.1)**
акција ‘action’	NOUN	агресија, акт, активност, атак, борба, чек, чин, дејство, делатност, дело, деловање, деоница, диверзија, иницијатива, инзулт, јуриш, кампања, корак, механизам, меница, нагласак, најезда, налет, напад, насиље, наскок, насртај, навала, облигација	деловање, поступак, операција, подухват, радња, мисија, интервенција, иницијатива, захват, покрет
бирати ‘choose’	VERB	изабирати, изабрати, одабрати, окренути, окретати, селектирати, требити	изабрати, одабрати, селектирати, одлучити, преферирати, опредељивати
акцијски ‘actional’	ADJ	акционарски, деоничарски, деонички	делатан, оперативан, динамичан, енергичан, подузетан, борбен, покретачки
данас ‘today’	ADV	сад, сада	сад, овогдана, овдашњи, овогтренутка, овогчаса

Table 1: Sample for synonyms support table; synonyms retrieved in lexical resources and suggested by GPT-4.1 are underlined.

each group. Here we will give a general prompt in English, while the original prompts were given in Serbian Cyrillic. Note that the similar approach was used for expressions, where supplementary lists were provided as google sheet documents.

```
system_instruction = """You are a lexicographer whose task is to write dictionary
    definitions in the Serbian language, respecting the language norm and style
    guidelines for dictionaries. The definition must not begin with the word being
    defined.
Bad example of a definition:
word: information; definition: Information is data about something or someone,
    notification; announcement, report.
Good example of a definition:
word: information; definition: data about something or someone, notification;
    announcement, report.
The definition must not contain the word being defined.
The definition must be one sentence.
Your answers shuold be in valid JSON format. The answer should contain only a list of
    entries with the fields "word", "sense number" and "definition".
Example answers:
[
{"word":"agronomist", "sense":"1", "definition":"a specialist in agronomy, agricultural
    engineer or agricultural technician; a student of an agricultural faculty."},
{"word":"gluttony", "sense":"1", "definition":"the quality of being greedy; gluttony."},
... ]
""")
```

A few examples illustrating this process are given in Table 2. One can observe that the number of meanings generated by Gemini-2.0 (the third column) and GPT-4.0 (the last column) differ, e.g. for 'закуцан' (nailed) only the basic meaning is given by GPT-4.0, while Gemini-2.0 lists three meanings, including one referring to sport (to dunk the ball). The adjective 'закуцан' (nailed) has no definition in Serbian dictionaries because its meaning is derived from the meaning of the corresponding verb 'to hammer'. The second meaning given by Gemini-2.0 'чврсто утврђен' (firmly established) is actually used, e.g. 'Курс евра је сада чврсто закуцан' (The euro exchange rate is now firmly established); however, this meaning does not appear in Serbian dictionaries.

Additional automation strategies include computing collocation frequencies, extracting typical syntactic patterns, and implementing GDEX algorithms for selecting illustrative usage examples directly from corpora, relying on NoSke[8], a local instance of the open-source NoSketchEngine platform (Kilgarriff et al., 2014, 2008), maintained by JeRTeh. NoSke supports numerous monolingual Serbian corpora, as well as bilingual corpora in which one of the languages is Serbian; however, for this dictionary, the most relevant resources are SrpKor2013 and SrpKor2021.

The underlying data model is inspired by and primarily aligned with the Data Model for Lexicography (DMLex) (Měchura, 2024), implemented using a PostgreSQL database. One of the modifications concerns the entry's headword, because DMLex only allows one while in our model there can be more: doublets and pronunciation variants (Ekavian and Ijekavian) can be added to the main headword. Another difference is that a headword in DMLex can have zero to more part of speech indicators associated with it, while in our model there is exactly one PoS per a headword (more is not supported).

4. RSSJ Application

The lexicographic workflow is supported by a custom-made web application. This user interface seamlessly integrates dictionary entry editing with corpus exploration via NoSke.

[8] https://noske.jerteh.rs/

Headword	PoS	Definitions (Gemini-2.0)	Definitions (GPT-4.0)
пумпање ‘pumping’	NOUN	1. радња пумпања; 2. убацивање или извлачење течности или гаса помоћу пумпе.	1. акција или процес премештања течности или гаса под притиском.
одређивати ‘determine’	VERB	1. утврђивати, прописивати; 2. постављати, налагати; 3. одлучивати, решавати.	1. одређивати постављати или утврђивати нечији положај, статус, природу или идентитет; 2. усмеравати или контролисати нечију активност или функцију.
закуцан ‘nailed’	ADJ	1. који је причвршћен куцањем; 2. чврсто утврђен, непоколебљив; 3. у спорту: који је постигао кош закуцавањем.	1. закуцан описује нешто што је чврсто причвршћено или закачено куком.
довека ‘forever’	ADV	1. довека за сва времена, заувек, вечно, непрекидно; 2. увек, стално.	1. довека синоним за ‘довек’, означава бескрајан, вечан временски период.

Table 2: A sample of generated definitions

Lexicographers can instantly query the corpora for concordances, collocations, or frequency information related to headwords, potential synonyms, or other lexical units directly from the editing environment using appropriate CQL queries. Figure 4 presents the main application panel for headword search (on the left side) and for entry editing (on the right side). The application can be used only by authorized users grouped in three roles: *editor*, who can write and edit selected entries; *redactor*, who can besides doing redaction also write, delete and edit all entries; and *administrator*, who can perform information dumping from the database using four different export templates.

Dictionary entries that are still in draft as well as those approved by the redactor can be displayed as HTML pages. An working version example is shown in Figure 5; note that the color scheme, style, and font can be easily adjusted as needed and multiple output styles can be defined.

Apart from the HTML format, the data export is available in *docx* format (MS Word document), with dictionary entries structured according to print requirements, including bold and italic formatting, while information is arranged according to the agreed dictionary entry specification. A sample is presented in Figure 6.

The third output option is used for a backup, and it produces a readable export of the database content in JSON format. The fourth export option is a SQL (relational) database dump. The initial part of the used script generates the database schema, while the second part populates the database with data.

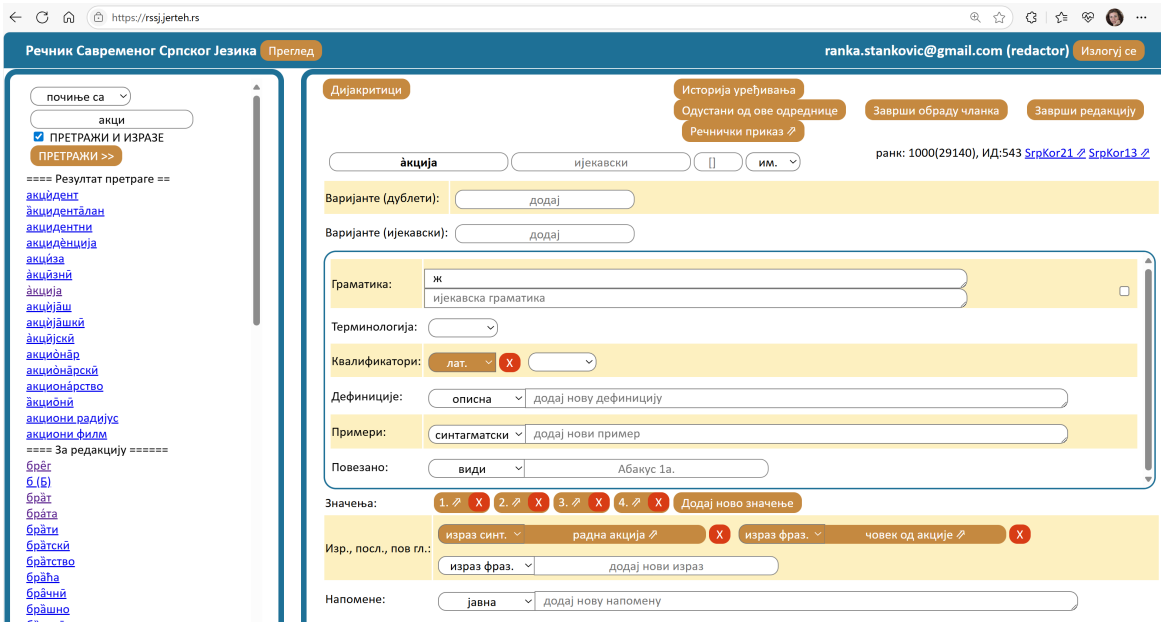

Figure 4: The main RSSJ application panel for data editing

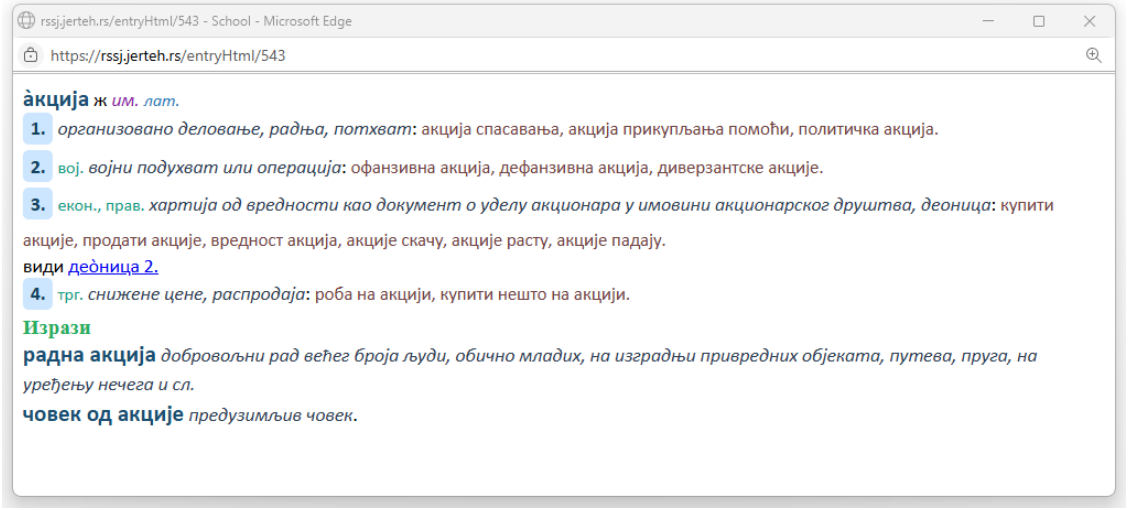

Figure 5: The draft of HTML entry preview

Once the dictionary is finalized, it will be available through separate browsing applications (web and mobile) with different levels of access. The general idea is to have the dictionary open for search, browse and API access: some of the API functions will be freely available while the other will be for authorized users only. Additionally, roughly 20% of dictionary entries will be available as open data for download.

5. Discussion

The experience gained through the RSSJ project so far underscores both the opportunities and ongoing challenges associated with digitally advanced lexicography for Serbian. Automation, driven by the integration of large corpora of contemporary Serbian (SrpKor2013, SrpKor2021), language models (Word2Vec, FastText, Dict2Vec), and LLMs (Gemini 2.0, ChatGPT 4.0), has led to substantial improvements in workflow efficiency, particularly in

àкција ж им. лат.
1. *организована друштвена или политичка делатност, организовано деловање, радња, потхват*: акција спасавања, акција прикупљања помоћи, хуманитарна акција, покренути акцију.
2. вој. *војна операција, војни подухват*: офанзивна акција, дефанзивна акција, диверзантске акције.
3. екон., прав. *хартија од вредности као документ о уделу акционара у имовини акционарског друштва, деоница*: купити акције, продати акције, вредност акција, акције скачу, акције расту, акције падају.
види деòница 2.
4. трг. *снижене цене, распродаја*: роба на акцији, купити нешто на акцији.
Изрази
радна акција *добровољни рад већег броја људи, обично младих, на изградњи привредних објеката, путева, пруга, на уређењу нечега и сл.*
човек од акције *предузимљив човек.*

Figure 6: An entry in the generated word document

the initial stages of candidate extraction and draft definition generation. The "blueprint" phase, an automatic compilation and structuring of up to 58,000 headwords, demonstrates a significant acceleration compared to traditional manual practices, enabling lexicographers to focus more on refinement and less on data collection.

However, the project also reveals several critical limitations and areas for further development. An important consideration is the coverage and balance of the lexicon itself. While automation makes it possible to include a broader spectrum of vocabulary and usage types, the representativeness and selection criteria must be carefully managed. The inclusion of diverse functional styles and the adaptation of the data model were crucial in addressing these issues, but further refinement will be necessary as the dictionary matures. Secondly, the quality of AI-generated definitions, while impressive for many entries, varies considerably, particularly for words with multiple meanings or complex semantic relationships. The experience of its use for Serbian lexicography, showed that content generated by artificial intelligence is not sufficiently reliable for lexicographers to depend on. For monosemous lexemes, especially terms named using international vocabulary, the representation of a given lexeme may be satisfactory. On the contrary, in the case of polysemous lexemes, numerous errors arise, ranging from incorrect sense disambiguation and definitions to inappropriate usage examples, grammatical errors, and orthographic mistakes. These shortcomings frequently result in misinterpretation of meanings (for example, the confusion between на памет (*to one's mind*) and напамет (*by hearth*, see the discussion below).

Here are some examples. The verb 'надвисивати' was defined as 'бити у високом положају изнад нечега, обично у претећем или доминантном положају' (*to be in a high position above something, usually in a threatening or dominant manner*) (ChatGPT), which does not correspond to the actual meaning of the lexeme. In addition to distinguishing its spatial meaning, **'**бити виши од некога или нечега**,** премашивати висином**'** (*to be taller than someone or something, to surpass in height*), from its figurative sense 'превазилазити, надмашивати' (*to surpass, to excel*), it is essential to emphasize that the verb always denotes excelling in positive attributes such as virtue, value, quality, significance,

ability, etc., and never implies a *'претећи или доминантни положај'* (*threatening or dominant position*). Similarly, in the definition of the verb 'навикавати' (*to get used to*) as 'стицати навику или привикавати некога на нешто' (*to acquire a habit or to accustom someone to something*) (Gemini), only the latter part: 'привикавати некога на нешто' (*to accustom someone to something*) is acceptable. The first part: 'стицати навику' (*to acquire a habit*) applies only to the reflexive form 'навикавати се'. In contrast to the perfective aspect of the verb 'навићи', which can carry the meaning 'стицати навику' (*to acquire a habit'*) (e.g., 'Нисам навикао да лежем рано' (*I am not used to going to bed early*), its imperfective aspect does not possess this meaning.

Recent artificial intelligence models provide some improved solutions, yet they remain far from satisfactory. For example, the dictionary entries, or at least some sections of the verb *'доћи'* (*to come*) and the preposition *'на'* (on, to) are presented fairly well. However, numerous errors persist: some meanings are illustrated using examples with the imperfective form of the verb 'долазити', e.g. 'Он долази сваког дана у исто време' *He comes every day at the same time*) or entirely ungrammatical constructions such as '*доћи да ради' (to come to work), '*доћи да каже' (to come to say).

Furthermore, distinct meanings are conflated under a single entry, as in the examples 'доћи до школе' (to come to the school), 'доћи до несреће' (an accident occurs), and 'доћи до договора' (to reach an agreement). In these cases, the verb *'доћи'* not only has a different meaning, but its usage is also not the same: in the second and third types of examples, unlike the first, it can only be used impersonally (e.g., 'Дошло је до несреће' (An accident has occurred), 'Не може доћи до договора' (An agreement cannot be reached). Similarly, The same meaning also includes examples such as 'доћи на власт' (to come to power) and 'доћи к себи' (to regain consciousness), even though these are separately listed under expressions ('доћи к себи') and phraseologisms ('доћи на власт').

In the processing of the preposition *'на'*, a lack of knowledge regarding orthographic rules became evident, leading to a cascade of additional errors. For instance, in the example '*Решио је задатке на памет' (He solved the tasks on mind), the phrase *'на памет'* (to one's mind) was used, whereas the correct form should have been the adverb *'напамет'* (by heart), formed by the fusion of the preposition *'на'* and the noun *'памет'* (*mind*). This adverb conveys the meaning 'не читајући, не гледајући у текст, без подсетника' (without reading, without looking at the text, without any prompts), which was the intended sense in this example (He solved the tasks by heart), a meaning not conveyed by the phrase *'на памет'*.

A similar error occurred with the phrase *'на мртво'*, which was incorrectly categorized as a phraseologism and, again erroneously, defined as 'без компромиса' (without compromise). In this context, the correct form is the adverb *'намртво'*, meaning 'to death' (e.g., 'пребити некога намртво' (to beat someone to death)).

Moreover, the preposition *'на'* was illustrated with ungrammatical examples, such as '*Радосна је на вест да ће постати бака' (She is happy at the news that she will become a grandmother). Other errors were also observed, many of which artificial intelligence can potentially correct through careful and persistent guidance. However, this requires additional time from the lexicographer, time which artificial intelligence is, in principle, intended to save.

All this leads to the conclusion that human review remains essential to ensure definitions are accurate, contextually appropriate, and stylistically aligned with dictionary standards. Gaps in sense coverage and missing meanings were observed, requiring manual intervention to supplement or correct automatically generated content. The introduction of AI and automated resources necessitates a shift in the role of lexicographers, from compilers and editors of manually created dictionary content to expert editors and evaluators of machine output. This shift requires new skills in prompt engineering, data validation, and corpus-based analysis. It also challenges lexicographers to develop more flexible, iterative workflows instead of following a strictly linear process.

The project's custom web application, seamless integration with corpus querying tools, and multi-format data exports (HTML, DOCX, JSON, SQL) illustrate the importance of infrastructure in supporting modern lexicographic practice. These tools not only enhance workflow efficiency but also facilitate collaboration, quality control, and future interoperability with external applications and research projects.

Finally, while the RSSJ project has exceeded its initial technical objectives, particularly in terms of automation and data management, a significant amount of lexicographic work remains. Completing and validating the full set of entries, enhancing the depth and accuracy of definitions, and addressing remaining gaps will require sustained human effort. The ongoing collaboration between computational and humane, lexicographic expertise remains central to the success of the project.

In summary, the RSSJ project highlights the potential for digital transformation in lexicography, illustrating that while automation and AI can significantly enhance efficiency and scalability, human expertise and editorial oversight are indispensable for ensuring quality, accuracy, and cultural relevance. The RSSJ application, along with its supporting tools and infrastructure, enables the continuous development and expansion of the dictionary by providing an integrated environment for editing entries, managing linguistic data, and incorporating new resources and user feedback on an ongoing basis. The first stable release of the dictionary is expected to be publicly available by the end of 2026.

6. Limitation

While the RSSJ project demonstrates substantial progress in automating dictionary development, several limitations remain:

- **Quality and consistency of AI-generated content.** Although large language models and embedding-based methods accelerate draft definition generation and candidate selection, the quality and consistency of AI-generated entries can be uneven. Automatically produced definitions may lack nuance, miss context, or fail to capture polysemy and culturally specific meanings, requiring continuous human oversight and manual correction.
- **Coverage gaps and missing senses.** Automated extraction is limited by the quality and scope of the underlying corpora and resources. Certain vocabulary, rare senses, or newly emerging usages may be underrepresented, necessitating additional manual curation and targeted expansion.

- **Editorial adaptation and skills gap.** The shift toward AI-assisted lexicography requires lexicographers to adapt to new workflows. This transition may present a skill gap, particularly for experts accustomed to traditional methodologies.
- **Evaluation and validation:** Systematic evaluation of AI-generated content remains an open challenge. Developing robust metrics and processes to assess linguistic accuracy, cultural appropriateness, and user relevance is necessary for ensuring the long-term quality and reliability of the dictionary.
- **Interoperability and standardization.** While the project leverages data models inspired by DMLex, further work is needed to ensure interoperability with lexicographic resources, platforms, and standards, particularly for open data publication and integration with external applications.

Addressing these limitations will require ongoing collaboration between computational and human expertise, IT experts, lexicographers and computational linguists, as well as the investment in the development of language resources, tools, and skills for the Serbian lexicographic community. The lessons learned from the automatic extraction of dictionary definition candidates from unstructured Serbian texts (Stanković et al., 2021) will be applied and evaluated using several definition templates.

7. Conclusion

The development of the Dictionary of Contemporary Serbian Language (RSSJ) demonstrates the transformative impact of advanced automation and AI-driven methodologies in modern lexicography. By integrating large corpora of contemporary Serbian, leveraging language models, and implementing robust data management strategies, the project contributed to accelerating the traditionally labor-intensive process of dictionary compilation. The custom-built application and flexible export formats further streamline editorial workflows and support a wide range of use cases for both human and machine consumers. Despite promising progress, challenges remain, particularly regarding the manual evaluation of automatically generated content and the ongoing need for expert lexicographic intervention. The preliminary results highlight that while AI-assisted drafting speeds up to a certain extent dictionary development, it cannot replace the nuanced judgment of human lexicographers. Future work will focus on refining automated approaches, enhancing coverage, and expanding interoperability with external resources and applications. Ultimately, the ongoing RSSJ project serves as a model for digitally advanced lexicographic initiatives and lays a solid foundation for the continued modernization of Serbian lexicography. The greatest contribution to the development of the RSSJ dictionary comes from linguistic corpora and the LeXimirka database, as they significantly accelerate and facilitate the compilation of dictionary entries. At present, the impact of large language models has not been substantial. However, we hope that future models, along with a synergy between traditional NLP techniques and language models, will yield improved results.

8. Acknowledgements

The project is supported by the effort “With Love”https://sljubavlju.com/ initiated and led by the Serbian diaspora. We would like to express our deep gratitude to the charitable association “Gathered around the Language” for launching and supporting the development of the Descriptive Dictionary of the Serbian Language.

Software

Abramski, K., Improta, R., Rossetti, G. & Stella, M. (2025). The "LLM World of Words" English free association norms generated by large language models. *Scientific data*, 12(1), p. 803.

Atkins, B.S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, pp. 135–146.

de Schryver, G.M. (2024). The Road towards Fine-Tuned LLMs for Lexicography. In *Workshop'Large Language Models and Lexicography'@ EURALEX 2024*. ELEXIS Association, pp. 6–11.

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. pp. 4171–4186.

Gantar, A. (2024). Formulating dictionary definitions using artificial intelligence using the example of Slovenian phraseological units [Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica]. In S. Marjanović (ed.) *Moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi*, volume 1 of *Leksikografski susreti*, chapter 12. Beograd: Univerzitet u Beogradu, Filološki fakultet, pp. 151–157. 12.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, pp. 7–36.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlỳ, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1. Universitat Pompeu Fabra Barcelona, pp. 425–432.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Klosa-Kückelhaus, A. & Tiberius, C. (2024). The Lexicographic Process Revisited. *International Journal of Lexicography*, 38(1), pp. 1–12. URL https://doi.org/10.1093/ijl/ecae016. https://academic.oup.com/ijl/article-pdf/38/1/1/60297676/ecae016.pdf.

Kosem, I., Krek, S. & Gantar, P. (2020). Defining collocation for Slovenian lexical resources. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8(2), pp. 1–27.

Kosem, I., Krek, S., Gantar, P., Holdt, Š.A. & Čibej, J. (2021). Language Monitor: Tracking the Use of Words in Contemporary Slovene. In *Electronic Lexicography in the 21st Century: Post-editing Lexicography. Proceedings of the eLex 2021 Conference*. p. 514. ELex 2021.

Krek, S. (ed.) (2024). *Large Language Models and Lexicography: Book of Abstracts of the Workshop*. Cavtat, Croatia: Centre for Language Resources and Technologies, University of Ljubljana & ELEXIS Association. Workshop Book of Abstracts.

Krstev, C. (2008). *Processing of Serbian – Automata, Text and Electronic Dictionaries*. Faculty of Philology, Belgrade.

Krstev, C., Pavlović-Lažetić, G. & Obradović, I. (2004). Using Textual and Lexical Resources in Developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, 7(1-2), pp. 147–161.

Lazić, B. & Škorić, M. (2019). From DELA-based dictionary to Leximirka lexical database. *INFOtheca: Journal of Information and Library Science*, 19, pp. 81–98.

Měchura, M. (2024). *Data Structures in Lexicography*. Ph.D. thesis, Ph. D. Thesis. Masaryk University. Brno A multilingual and multifunctional ….

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Navigli, R. & Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. pp. 1318–1327.

Pilehvar, M.T. & Camacho-Collados, J. (2021). Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. *Computational Linguistics*, 47(3), pp. 699–701.

Stanković, R., Krstev, C., Lazić, B. & Škorić, M. (2018a). Electronic dictionaries–from file system to lemon based lexical database. In *6th Workshop on Linked Data in Linguistic (LDL-2018), Towards Linguistic Data Science*. pp. 48–56.

Stanković, R., Krstev, C., Stijović, R., Škorić, M. & Gočanin, M. (2021). Towards automatic definition extraction for Serbian. In *EURALEX (Proceedings of the XIX EURALEX Congress of the European Assocition for Lexicography: Lexicography for Inclusion, Vol. 2*. Komotini : SynMorPhoSe Lab, Democritus University of Thrace, pp. 695–704.

Stanković, R., Mladenović, M., Obradović, I., Vitas, M. & Krstev, C. (2018b). Resource-based WordNet augmentation and enrichment. In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*. pp. 104–114.

Stanković, R., Rađenović, J., Škorić, M. & Putniković, M. (2025). Learning Word Embeddings using Lexical Resources and Corpora. In *Proccedings of 15th International Conference on Information Society and Technology – ICIST 2025*.

Stijović, R., Krstev, C. & Stanković, R. (2021). Automatska ekstrakcija definicija – doprinos ubrzanju izrade rečnika / Automatic Definition Extraction – Accelerating Dictionary Development. In *Leksikologija i leksikografija u svetlu aktuelnih problema / Lexicology and Lexicography in the Light of Current Issues*. Beograd: Institut za srpski jezik SANU, pp. 113–137.

Stijović, R., Stanković, R. & Škorić, M. (2025). Dictionary of Modern Serbian Language: RSSJ. In *Book of Abstracts of the International Conference South Slavic Languages in the Digital Environment - JuDig*. p. 32.

Tiberius, C., Kallas, J., Koeva, S., Langemets, M. & Kosem, I. (2024). A Lexicographic Practice Map of Europe. *International Journal of Lexicography*, 37(1), pp. 1–28.

Tufis, D., Cristea, D. & Stamou, S. (2004). BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology*, 7(1-2), pp. 9–43.

Vitas, D. & Krstev, C. (2012). Processing of corpora of serbian using electronic dictionaries. *Prace Filologiczne*, LXIII, pp. 279–292.

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

Vitas, D., Stanković, R. & Krstev, C. (2025). The Many Faces of SrpKor. In *Proceedings of the International Conference South Slavic Languages in the Digital Environment - JuDig*. pp. 1–28.