
Речник савременог српског језика: од дигиталне лексикографске базе до отворене језичке инфраструктуре

Ранка Станковић¹[0000-0001-5123-6273], Михаило
Шкорић¹[0000-0003-4811-8692], Рада Стијовић¹

¹ Друштво за језичке ресурсе и технологије – ЈеРТех, Србија
{ranka|mihailo}@jerteh.rs, stijovicr@gmail.com

Научни рад

DOI: 10.18485/judig.2026.2.ch4

Апстракт

Речник савременог српског језика (РССЈ) представља лексикографски подухват чији је циљ изградња савременог речника српског језика заснованог на корпусним подацима, савременим лексикографским принципима и језичким технологијама. За разлику од традиционалних речника, РССЈ је замишљен као интегрисана дигитална инфраструктура која обједињује речничку базу података, лексикографско окружење, корпусне ресурсе, програмске интерфејсе, веб и мобилне апликације. Рад представља архитектуру система и модел података имплементиран у систему PostgreSQL, као и интеграцију са корпусом СрпКор2021+ на платформи NoSketch Engine за претрагу корпуса. Посебно се разматра концепт програмског интерфејса REST API, који омогућава да речничким подацима приступају и људи и машине, чиме се речник повезује и са другим језичким ресурсима, попут лексичке базе Лексимирика и Српског ворднета. Пажња је посвећена и веб и мобилним апликацијама, механизмима аутентификације и управљања приступом, као и могућностима интеграције са системима заснованим на вештачкој интелигенцији. Рад показује како се традиционална лексикографија може трансформисати кроз савремену дигиталну инфраструктуру способну да подржи истраживачки рад, образовање, индустријске примене и развој језичких технологија за српски језик.

Кључне речи: дигитална лексикографија, речничке базе података, API, веб-апликације, мобилне апликације, језичке технологије, српски језик

1. Увод

Савремени речници више нису само публикације намењене човеку, већ представљају комплексне дигиталне инфраструктуре које омогућавају приступ лексичком знању путем веб и програмских интерфејса, мобилних апликација и система вештачке интелигенције. Последње две деценије обележене су преласком са традиционалне папирне лексикографије на дигиталне лексичке ресурсе, који омогућавају континуирано ажурирање, аутоматизовану обраду и поновну употребу података у различитим информационим системима. (Klosa-Kückelhaus and Tiberius, 2024; Tiberius et al., 2024)

Овај прелаз, међутим, не доноси само нове могућности већ и нове изазове. Лексички ресурси су често развијани независно, у међусобно неусаглашеним форматима, што отежава њихову размену, поновну употребу и повезивање са алатима за обраду природних језика. Речници су, услед тога, неретко остајали изоловани од ширег екосистема језичких технологија, а подаци похрањени у њима тешко доступни за машинску обраду. Управо превазилажење ове фрагментисаности — кроз стандардизован модел података и програмски приступ речничким ресурсима и корпусима представља једну од кључних мотивација рада који следи.

У том контексту развија се и Речник савременог српског језика (РССЈ), чији је циљ израда око 50.000 речничких чланака који описују лексику савременог српског стандардног језика (Стијовић et al., 2024; Stanković et al., 2025). За разлику од класичног речника, РССЈ је конципиран као централна лексикографска платформа која повезује корпусне ресурсе, лексичке базе података, језичке моделе и софтверске компоненте за приступ речничким подацима и њихово коришћење (Vitas and Krstev, 2015, 2012). Развој речника ослања се на корпус СрпКор2021+ (Витас et al., 2025), који обједињује корпусе СрпКор2013 и СрпКор2021, лексичку базу Лексимирка (Stanković et al., 2018a; Lazić and Škorić, 2019), Српски ворднет (Krstev et al., 2004; Stanković et al., 2018b) и друге језичке ресурсе (Krstev, 2008).

Основни циљ рада је представљање архитектуре овог интегралног система, интеракције са корпусом и нових компоненти система: програмског интерфејса (Application Programming Interface, API), веб-апликације и мобилне апликације. Речничка база је имплементирана у систем за управљање релационим базама пода-

така PostgreSQL¹. Модел података је у великој мери усаглашен са стандардом DMLex (Data Model for Lexicography)², који је стандардизовани модел за структурирање речника и лексикографских ресурса на флексибилан начин, независан од конкретне технологије. Поред лексикографског окружења за израду речничких чланака, развијене су и нове компоненте намењене дистрибуцији и експлоатацији речничких података: програмски интерфејс (API), веб-апликација и мобилна апликација.

1.1. Дигитална лексикографија и стандарди за представљање података

Савремена дигитална лексикографија заснива се на структурираним моделима података као што су DMLex, TEI Lex-0³ и OntoLex-Lemon. Ови стандарди омогућавају интероперабилност лексичких ресурса и њихову интеграцију у шири екосистем језичких технологија.

DMLex је апстрактни модел података, а не формат записа. Прво се дефинише апстрактна структура, која се потом „преводи” у конкретне формате (XML, JSON, релациону базу података, RDF и др.). У раду је примењен централни модул стандарда (DMLex Core) заједно са модулом за повезивање, док остали модули нису искоришћени у пуном обиму. Модел података РССЈ обезбеђује јасно раздвајање леме, њених значења, граматичких информација и примера употребе, што омогућава једноставну размену података између различитих речничких система.

По приступу изградњи дигиталне лексикографске инфраструктуре, РССЈ се уклапа у ток савремених националних пројеката дигиталних речника у Европи, попут словеначких, чешких и холандских иницијатива за изградњу отворених и интероперабилних лексичких ресурса (Krek et al., 2025). На тај начин, развој речника не прати само домаћу лексикографску традицију већ и међународне стандарде и добре праксе у области лексикографије (Tiberius et al., 2024).

¹ <https://www.postgresql.org/>

² <https://docs.oasis-open.org/lexidma/dmllex/v1.0/dmllex-v1.0.pdf>

³ <https://lex0.org/>

1.2. Инфраструктурни пројекти и вештачка интелигенција у лексикографији

Међу инфраструктурним иницијативама у области дигиталне лексикографије посебно се издваја пројекат ELEXAI (European Lexicographic Infrastructure for Artificial Intelligence). ELEXAI је пројекат финансиран у оквиру програма Хоризонт, чији је циљ, с једне стране, унапређење постојећих модела вештачке интелигенције, а с друге, темељна надоградња Европске лексикографске инфраструктуре (ELEXIS), настале у истоименом пројекту Хоризонт 2020 (2018–2022)⁴. Надоградња се заснива на најновијим достигнућима вештачке интелигенције, пре свега на појави великих језичких модела. Осим интеграције националних, регионалних и институционалних лексикографских иницијатива, нова инфраструктура нуди унапређене технологије, језичке податке, алате и сервисе намењене побољшању трансформерских модела квалитетним лексикографским знањем. Идеја је да се лингвистички утемељена машински читљива репрезентација знања уграђује у језичке моделе, чиме се добијају бољи резултати, који се потом користе за даље усавршавање самих језичких ресурса.

Последњих година се овакви системи све чешће користе за аутоматско генерисање дефиниција, избор добрих речничких примера (Good Dictionary EXamples, GDEX) (Kilgariff et al., 2008), семантичку класификацију, аутоматску екстракцију лексикографских информација (Kosem et al., 2020) и праћење језичких промена у савременом језику (Kosem et al., 2021).

1.3. Корпусно заснована лексикографија

Још један важан аспект је лексикографија заснована на великим текстуалним корпусима. Модерни речници све више користе такве корпусе као основни извор информација о значењу, контексту и обрасцима појављивања речи. Корпусна лингвистика омогућава екстракцију колокација, фреквенција, типичних образаца употребе и примера који могу бити укључени у речничке чланке. Управо се на овај начин корпус СрпКор2021+ користи као емпиријска основа за израду и допуну речничких чланака у оквиру РССЈ.

⁴ <https://elex.is/>

1.4. Организација рада

Рад је организован на следећи начин. У одељку 2. приказују се речничка база и лексикографско окружење, и представљање опште архитектуре система и његових главних компоненти у одељку 2.1. Одељак 2.2. описује модел података и његову усаглашеност са стандардом DMLex. У одељку 3. детаљно се разматрају програмски интерфејс и клијентске апликације (веб у одељку 3.3. и мобилна у одељку 3.4.), уз примере њихове употребе. У одељку 4. разматрају се постигнути резултати, ограничења предложеног приступа и изазови везани за интеграцију речничких података и клијентских апликација. Одељак 5. доноси закључна разматрања и правце даљег рада.

2. Речничка база и лексикографско окружење

Основу система чини речничка база података која садржи поља за унос одредница, граматичких података, дефиниција, примера употребе, израза, пословица, синонима и упућивања. Одреднице су иницијално генерисане аутоматским ексцерпирanjem из корпуса по фреквенцијама, а затим допуњене подацима из постојећих речника и лексичких ресурса.

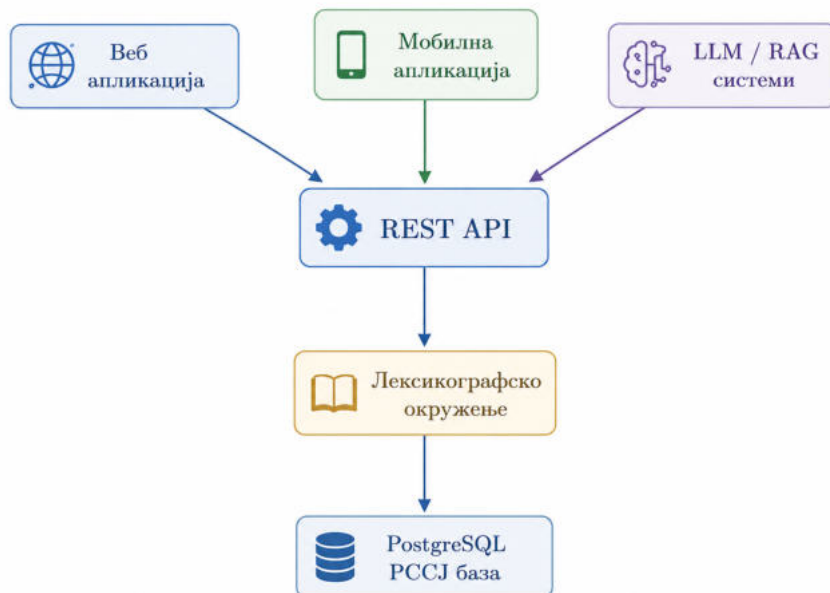
Лексикографима је на располагању веб окружење које интегрише уређивање речничких чланака са директним приступом корпусима преко локалне инстанце платформе NoSketch Engine (Kilgarriff et al., 2004, 2014) и корпуса *СрпКор2013* и *СрпКор2021* (Витас et al., 2025), касније спојених у *СрпКор2021+*. На тај начин је омогућено претраживање конкорданци, колокација, фреквенција и других корпусних показатеља непосредно током израде речничког чланка.

2.1. Архитектура система

Систем РССЈ организован је као вишеслојна архитектура која раздваја складиштење података, лексикографску обраду, програмски приступ и корисничке интерфејсе (слика 1). Оваква организација омогућава независан развој појединачних компоненти, лакше одржавање система и једноставну интеграцију са спољашњим апликацијама.

У центру архитектуре налази се речничка база података реализована у систему PostgreSQL, који представља софтвер за

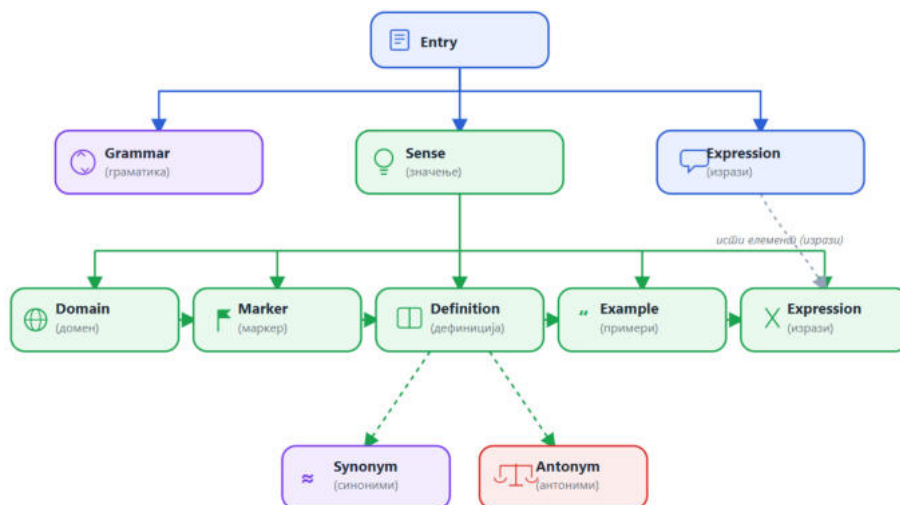
управљање базама података, односно програм који организује, чува и претражује велике количине структурираних података. У њему су речнички чланци смештени у прегледне табеле са јасно дефинисаним пољима (нпр. одредница, врста речи, дефиниција, примери), што омогућава брзо и поуздано претраживање. Изнад базе се налази лексикографско окружење намењено уређивању и одржавању речничких чланака. Приступ подацима омогућен је преко REST API (Representational State Transfer API) слоја. Реч је о посредничком слоју који стоји између базе и програма који је користе: уместо да свака апликација директно приступа бази, она своје захтеве шаље API-ју у унапред договореном облику (нпр. „врати ми све податке о речи вода”), а API враћа одговор у стандардизованом формату погодном за даљу обраду. На тај начин се обезбеђује једнообразна и контролисана комуникација између базе и различитих клијентских апликација, независно од тога на којој су технологији изграђене. Највиши слој чине веб-апликација, мобилна апликација и системи вештачке интелигенције (LLM/RAG) који користе речничке податке.



Слика 1: Вишеслојна архитектура система РССЈ.

2.2. Модел података

Модел података (слика 2) инспирисан је стандардом DMLex, али је прилагођен специфичностима српског језика. Основни ентитет представља заглавље речничког чланка (**Entry**), које је повезано са значењима, примерима употребе, синонимима, изразима, пословицама и различитим квалификаторима.



Слика 2: Поједностављени UML модел података РССЈ

Основна јединица модела јесте одредница (**Entry**), из које се гранају три главна елемента: граматика (**Grammar**), изрази (**Expression**) и значења (**Sense**). Значење представља најразвијенији део структуре и обухвата више целина: домене употребе (**Domain**) и маркере (**Marker**), затим ентитет за дефиниције (**Definition**), које могу бити описне, синонимске или упућивачке, потом примере (**Example**) — синтагматске и реченичне. На крају су изрази и пословице (**Expression**). Додатно се успостављају релације ка сродним значењима речи, синоними (**Synonym**), и ка супротним значењима, антоними (**Antonym**), који заједно чине лексичко-семантичке односе између значења.

3. Програмски интерфејс за приступ речнику

Један од кључних циљева пројекта је омогућавање машинског приступа речничким подацима. У ту сврху развија се

интерфејс типа REST API, који за размену података користи JSON формат.

Интерфејс (API) омогућава приступ различитим нивоима речничке структуре, од једноставног претраживања лема до добијања комплетних речничких чланака. Предвиђене су функције за:

- претрагу речи према обрасцима (почиње са, завршава се на, садржи, слична је);
- излиставање синонима и антонима;
- излиставање врста речи, дефиниција, домена употребе и квалификатора;
- претраживање по доменима употребе и квалификаторима;
- прегледање екавских и ијекавских облика;
- приказивање комплетног речничког чланка према идентификатору (у JSON или HTML облику);
- приказивање комплетног речничког чланка на латиници.

Поред основног скупа функција, у каснијим фазама развоја планира се увођење следећих додатних сервиса:

- проналажење семантички сличних речи на основу векторских репрезентација;
- статистика коришћења API функција;
- API кључеви и управљање квотама приступа;
- праћење различитих верзија речничких чланака;
- претрага по деловима дефиниције;
- извоз података у RDF/JSON-LD формату;
- OpenAPI (Swagger) документација.

Посебна пажња посвећена је управљању правима приступа. Предвиђени су нивои јавног, регистрованог и комерцијалног приступа, док ће део речничке грађе бити доступан у виду отворених података.

3.1. Интеракција између речника и корпуса

Повезивање апликације за речник са корпусом омогућава да се конкорданце добијају директно из речничког окружења — уместо да корисник ручно формулише упит у корпусном претраживачу, апликација на основу саме одреднице аутоматски генерише одговарајући упит и приказује потврде из корпуса. За једночлане одреднице тај поступак је једноставан: из леме и граматичке ознаке (врсте речи) генерише се непосредан упит облика

[lemma=". . ." & tag=". . ."], којим се обухватају сви облици те речи у корпусу. Интеграција апликације и корпуса омогућава и приказивање колокација за унапред дефинисане синтаксичке обрасце.

Код вишечланих одредница, међутим, поступак је знатно сложенији, јер сваки члан израза мора да се пресликава у засебан токен, уз навођење свих релевантних лема и, по потреби, ограничења по врсти речи, а њихов међусобни редослед и растојање морају се очувати. Додатно усложњавање настаје када израз садржи заграде: пошто оне сигнализирају необавезне или променљиве сегменте, генератор CQL упита не сме да их третира као обичан текст, већ мора да их препозна и преведе у одговарајуће CQL конструкције — необавезне групе (. . .)? за оно што може, али не мора да се појави, односно празне токене [] за променљиве допуне — чиме се аутоматско генерисање упита из формалног пресликавања претвара у поступак који захтева разлагање структуре израза и различит третман сваке врсте заграда.

У изворним речничким одредницама традиционалних описних речника српског језика, све што је необавезно или променљиво обележено је истом врстом заграда — обликом (). Проблем је што те заграде у речнику покривају три суштински различите језичке појаве, а свака се преводи на потпуно другачији начин. Због тога једна врста заграда није довољна: систем не може да разазна шта се тачно тражи и прави грешке.

Решење уведено у РССЈ јесте увођење три врсте заграда, од којих свака недвосмислено сигнализира начин на који се дати сегмент интерпретира и аутоматски преводи у упит над корпусом.

Обле заграде су довољне само када обележавају опциони део речи. Пошто се ту не додаје нити одузима засебна реч, ствар се решава простим набрајањем лема у оквиру једне угласте заграде. На пример: ‘(на)правити се луд’ ће се трансформисати у упит: [lemma="napraviti|praviti"] [lemma="se"] [lemma="lud"].

У овом случају заграда обухвата морфему, а не реч, па се обе варијанте (са делом и без њега) наводе као алтернативне леме унутар истог токена, тако да додатне врсте заграда нису потребне.

У примеру: ‘делити (држати) лекцију’ заграда не означава ни необавезни, ни променљиви члан, него алтернативну реч на истом месту: „делити“ и „држати“ су заменљиви глаголи који попуњавају исти слот у изразу („делити/држати лекцију“). Пошто

обе варијанте стоје на истој позицији (један токен), не праве се ни необавезна група (. . .)? ни празан токен [], већ се обе леме једноставно наводе као алтернативе унутар истог токена, раздвојене усправном цртом: [lemma="deliti|držati"] [lemma="lekcija"]. Приметимо да је систем од облика „лекцију“ препознао лему „лекција“. У овој верзији се задржавамо на лемама, не радимо синтаксичку предикцију израза (који је променљиви, а који непроменљиви део), већ све токене представљамо лемом (или са више лема када је вишезначно). Циљ нам је што већи одзив, под цену и да прецизност буде нешто нижа.

Поменимо још један случај чест у речницима који захтева додатну интерпретацију. На пример израз: ‘на својим леђима вући (држати, имати, носити и сл.)’ треба на неки начин проширити да обухвати сличне изразе. Један начин би био да се листа прошири синонимима, али смо се ми и у овој имплементацији одлучили за решење које повећава одзив по цену прецизности.

Наиме, систем провери која је врста речи у листи која се набраја, а затим дода одговарајући токен са рестрикцијом на врсту речи, тако да у овом случају добијамо упит: [lemma="na"] [lemma="svoj"] [lemma="leđa|leđima"] ([lemma="vući"] | [lemma="držati"] | [lemma="imati"] | [lemma="nositi"] | [tag="V"]) који ће препознати у корпусу конкорданце са следећим глаголима: носити (48), изнети (30), осетити (14), понети (5), вући (4), осећати (4).

Проблем настаје када обла заграда у изразу обухвата целу реч, а не део речи. Решили смо га увођењем витичастих заграда { } којима се обележава да је цела реч необавезна, или да је више речи необавезно. Дакле, део израза који може, али не мора да се појави, у упиту се преводи као засебна, необавезна група (. . .)?. На пример, уместо традиционалног записа ‘(као) за баксуз’, у РССЈ пишемо ‘{као} за баксуз’, што се преводи у упит ([lemma="kao"])? [lemma="za"] [lemma="baksuz"].

Витичасте заграде природно покривају и необавезан низ алтернатива (спој необавезности и избора преко |). На пример: ‘бранити боје {клуба, репрезентације, земље}’ трансформисаће се у упит: [lemma="braniti"] [lemma="boj|boja|bojati|bojiti"] ([lemma="klub"] | [lemma="reprezentacija"] | [lemma="zemlja"])?.

Угласе заграде [] представљају променљиву (било која реч), односно слободно место које може попунити било која реч (обично допуна: „некога“, „нечију“). У упиту се преводи као празна угласа заграда [], која се поклапа с било којим појединачним токеном. Традиционално записани израз: ‘бинути (нечију) бригу’ сада записујемо као ‘бинути [нечију] бригу’ што се преводи у упит: [lemma="brinuti"] [tag_u="PRON|DET"] [lemma="briga"]. Празна заграда [] управо решава овај случај — прихвата сваку реч на том месту.

Разлика се најбоље види на изразу који садржи и варијантни део речи, и факултативну реч, и променљиву допуну. Замислимо одредницу: ‘(про)читати {пажљиво} [нешто]’. Префикс ‘(про)’ представља варијантни део тако да се у упиту трансформише у [lemma="pročitati|čitati"]. Реч ‘{пажљиво}’ је факултативна, тако да се користе витичасте заграде и преводи у упит као [lemma="pažljivo"]?, док је допуна [нешто] је променљива допуна за коју користимо угласе заграде, и то се у упиту преводи у [‘’]. Дакле, све заједно: [lemma="pročitati|čitati"] ([lemma="pažljivo"])? [‘’]

Да су сва три сегмента остала у обим заградама, систем не би могао да разлучи да ли „пажљиво“ факултативно, какав статус има „нешто“, тако да управо ту разлику три врсте заграда чине експлицитном. Табела 1 представља сажети преглед значења заграда у предложеној нотацији.

Табела 1: Преглед значења заграда у нотацији коришћеној за изразе.

Заграде	Значење у изразу	Превод у упит
()	необавезан <i>део речи</i> (морфема)	набрајање лема у истом токenu: [lemma="napraviti praviti"]
{ }	необавезна <i>цела реч / израз</i>	необавезна група: (...)?
[]	<i>променљива</i> (било која реч)	празан токен: []

Анализирано је укупно 137 израза. Од тог броја, 74 израза (54%) остало је непромењено у односу на традиционални приступ, док је 63 израза (46%) промењено мењањем облика заграда у угласте и витичасте. У 12 случајева (9%) претрага није вратила резултате (3 случаја из скупа непромењених израза и 9 из скупа промењених). Од ових 12 случајева, у корпусу нису нађене конкорданце за 8 (6%) израза, а у 4 случаја (3%) реч је о нерешеној синтакси.

Сам процес генерисања упита невидљив је за корисника: када се, позициониран на запису са изразом, кликне на везу ка корпусу, систем генерише одговарајући упит и прослеђује га, а корисник добија конкорданце. По потреби, корисник може и да измени упит уколико добија преширок одзив.

Сличан механизам генерисања упита искоришћен је и за аутоматско издвајање колокација. За сваку одредницу, на основу њене врсте речи, припремљени су одговарајући граматички обрасци који описују типичне колокацијске позиције у односу на кључну реч. Тако се, на пример, за именицу издвајају придеви који јој претходе (нпр. тежак проблем), именице у непосредном окружењу (проблем решења, решавање проблема), као и глаголи с којима се често јавља (решити проблем, наметнути проблем), док се за глагол издвајају типични субјекти и објекти, и слично. Сваки образац преводи се у одговарајући CQL упит који, уз ограничење по врсти речи и растојању од кључне речи, у корпусу проналази све потврде дате колокацијске везе.

Резултати добијени упитом рангирају се по фреквенцији (и, по потреби, по статистичким мерама асоцијације попут *logDice* или узајамне информације), чиме се лексикографу непосредно током израде чланка нуди преглед најчешћих и најтипичнијих колокација дате речи. На тај начин колокације престају да буду плод ручног прегледа конкорданци и постају аутоматски понуђен, емпиријски утемељен материјал који лексикограф само проверава и бира — што је у складу са општим приступом система да претрагу корпуса приближи што више самом чину израде речничког чланка.

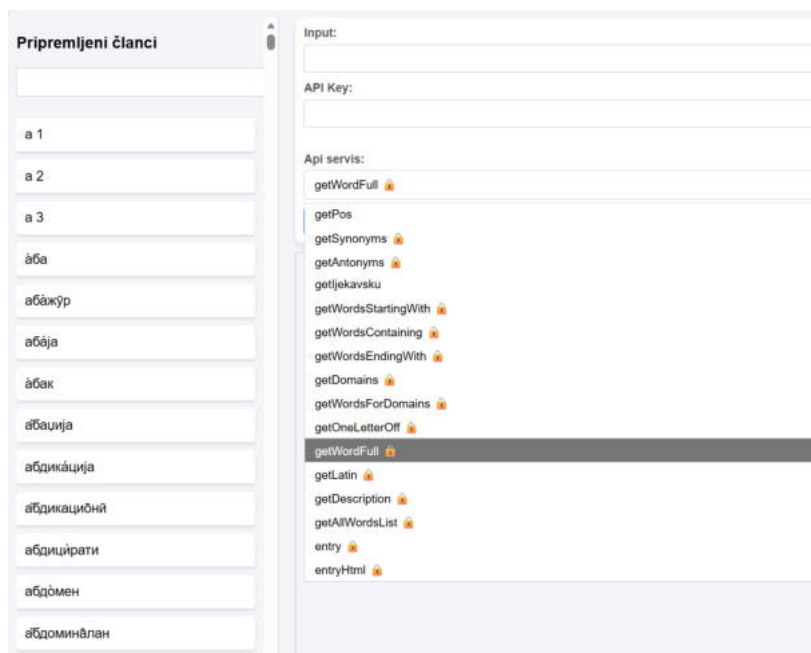
3.2. API функције

Основни скуп API функција⁵ приказан је у Табели 2, а приказ API интерфејса приказан је на Слици 3.

Табела 2: Основне API функције, формати улаза и излаза, као и опис сваког од излаза.

Функција	Опис излаза	Улаз	Излаз
getWordsStartingWith	Листа речи које почињу са	текст	JSON
getWordsContaining	Листа речи које садрже	текст	JSON
getWordsEndingWith	Листа речи које се завршавају са	текст	JSON
getOneLetterOff	Листа речи које су сличне као	текст	JSON
getWordFull	Комплетан речнички чланак за задату реч	текст	JSON
getLatin	Комплетан речнички чланак на латиници за задату реч	текст	JSON
entry	Комплетан речнички чланак за задати ID	број	JSON
entryHtml	Комплетан речнички чланак за задати ID (HTML)	број	HTML
getPos	Врста речи за задату реч	текст	JSON
getDescription	Листа дефиниција за задату реч	текст	JSON
getIjekavsku	Ијекавска варијанта за задату реч (ако постоји)	текст	JSON
getSynonyms	Листа синонима за задату реч	текст	JSON
getAntonyms	Листа антонима за задату реч	текст	JSON
getDomains	Листа домена/квалификатора за задату реч	текст	JSON
getWordsForDomain	Листа речи за задати домен	текст	JSON
getAllWordsList	Комплетна листа речи		JSON

⁵ <https://rssj.jerteh.rs/api>



Слика 3: Интерфејс за увид и анализу API функција система РССЈ.

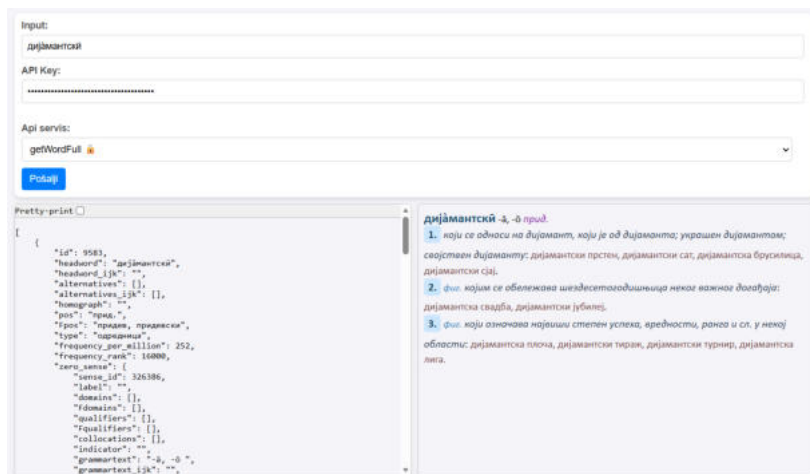
Као пример излаза издвојена је функција `getWordFull`, која враћа све податке похрањене за један речнички чланак и захтева API кључ. Њен излаз приказан је на слици 4. Корисник претражи речи, изабере одговарајућу и унесе кључ; када позове функцију, у доњем делу екрана са леве стране приказује се JSON одговор намењен машинском читању, а са десне стране облик разумљив човеку.

На слици 5 приказан је пример јавно доступне функције која не захтева кључ — функција `getIjekavsku` за задату реч „дете“ враћа њен ијекавски еквивалент „дијете“.

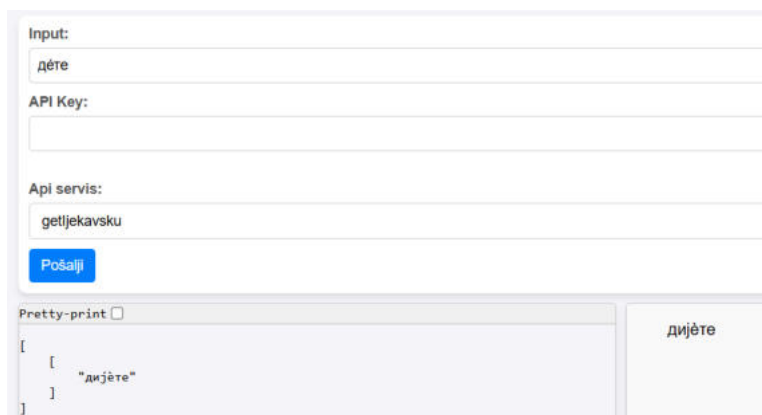
3.3. Веб-апликација

Веб-апликација представља примарни начин приступа речнику за већину корисника. Основне функције обухватају:

- једноставну и напредну претрагу;
- преглед комплетног речничког чланка;
- навигацију кроз сродне појмове;
- приказ израза, пословица и синонима;



Слика 4: Пример одговора функције `getWordFull` за реч *дијамантски*.



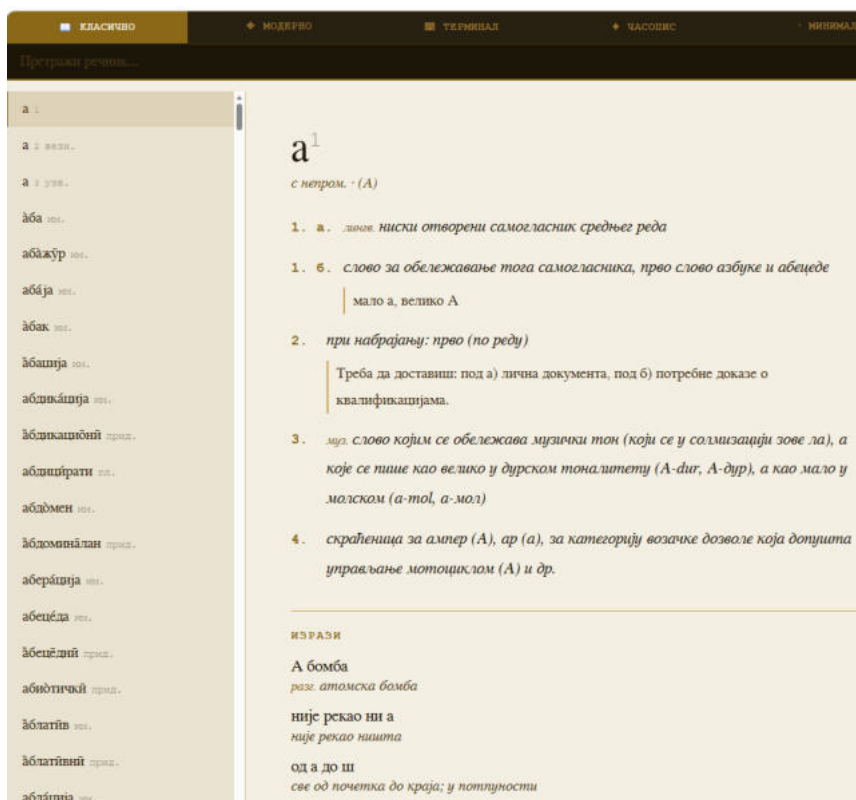
Слика 5: Пример функције `getIjekavsku` за реч *дете*.

- преглед акценатских дублета и варијаната (екавских и ијекавских);
- дељење и цитирање речничких чланака.

Са становишта корисничког искуства, веб-интерфејс (слика 6) треба да подржи светлу и тамну тему, респонзивни приказ⁶ на различитим уређајима, брзо претраживање и оптимизован приказ сложених речничких чланака.

⁶ Респонзивни приказ омогућава прилагођавање садржаја различитим величинама екрана.

Поред лексикографског садржаја, апликација треба да садржи информације о пројекту, институцијама које учествују у његовом развоју, као и механизме за донације и подршку пројекту.

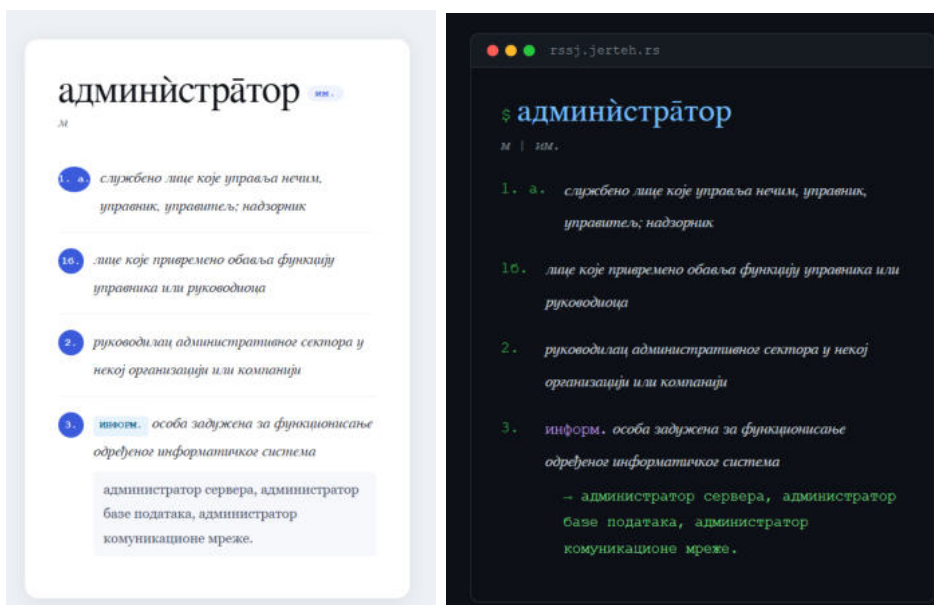


Слика 6: Панел веб-апликације

3.4. Мобилна апликација

Мобилна апликација (слика 7) омогућиће коришћење речника на Android уређајима, а касније и на iOS платформи. Основне функције мобилне апликације обухватају:

- брзу претрагу речи;
- приказ комплетног речничког чланка;
- листу недавно претраживаних речи;
- листу омиљених одредница;
- дељење садржаја путем других апликација;
- рад у светлом и тамном режиму.



Слика 7: Панел мобилне апликације (светла и тамна верзија)

У наредним фазама развоја планира се и додавање следећих функција: офлајн режим рада, персонализоване листе речи, реч дана, едукативни садржај и интеграције са језичким технологијама заснованим на вештачкој интелигенцији. Офлајн режим омогућио би коришћење речника и без интернет везе, локалним чувањем података на уређају. Интеграција са вештачком интелигенцијом односила би се, на пример, на повезивање речника са великим језичким моделима преко RAG приступа, у коме модел преко API-ја преузима речничке податке и на њима заснива тачнији одговор.

4. Дискусија

Један од најзначајнијих праваца даљег развоја јесте интеграција великих језичких модела у процес израде речничких чланака. Аутоматизација је увела тзв. модел пост-едиторске лексикографије, у коме се улога лексикографа помера ка провери и доради аутоматски генерисаних нацрта речничких чланака (Marković and Stanković, 2025). Могућности примене оваквих модела у српској дескриптивној лексикографији — генерисању дефиниција, разрешавању вишезначности вишечланих израза и креирању илу-

стративних примера — испитане су на скупу од око 400 идиоматских вишечланих израза, при чему су поређени резултати више савремених модела (ChatGPT-4o и 4.1, Gemini-2.5-Flash и 2.5-Pro). Резултати оваквих истраживања представљају основу за будуће повезивање речничке инфраструктуре РССЈ са системима заснованим на вештачкој интелигенцији.

Лексички ресурси представљају поуздан извор висококвалитетног, људски провереног лингвистичког знања које може значајно допринети развоју и унапређењу великих језичких модела. За разлику од сирових веб-корпуса, лексикографски ресурси су брижљиво састављени, семантички структурисани и лингвистички валидирани, што их чини посебно значајним за језике са ограниченим ресурсима, какав је и српски. Њихова интеграција у системе вештачке интелигенције отвара више важних истраживачких праваца у области могуће примене лексичких ресурса у контексту великих језичких модела:

- генерисање синтетичких текстова који се могу користити за предобучавање специјализованих модела;
- генерисање задатака за праћење инструкција ради побољшања способности разумевања и закључивања модела;
- изградњу графова знања заснованих на лексичким и семантичким ресурсима ради подршке апликацијама;
- интеграцију са системима заснованим на приступу Retrieval-Augmented Generation (RAG), кроз преузимање речничких одредница, подграфова знања и SQL упита;
- развој агената заснованих на великим језичким моделима који могу аутоматски писати код и издвајати корисне информације из хетерогених лексичких и језичких ресурса;
- креирање изазовних тестова за евалуацију лингвистичке компетенције, семантичког разумевања и способности закључивања великих језичких модела.

Ови приступи показују да лексикографија више није ограничена искључиво на традиционалну израду речника, већ постаје важна компонента у развоју нове генерације језичких технологија. Посебно је значајно то што структурисано лексичко знање може допринети већој фактографској поузданости, интерпретабилности и доменској адаптацији великих језичких модела, истовремено омогућавајући богатије семантичко утемељење и контролисаније генерисање текста.

Речнички примери могу послужити и као основа за развој језичких модела у задацима разрешавања и индукције значења речи, што је посебно значајно за језике са ограниченим ресурсима за ове задатке, попут српског. Истраживања на словеначком језику показала су да се управо из речничких примера, уз помоћ великих језичких модела, могу успешно генерисати подаци за обуку таквих модела, чиме речник постаје вредан извор за унапређење језичких технологија ([Škvorc and Robnik-Šikonja, 2025](#)).

Један од стратешких праваца развоја PCCJ јесте његова интеграција са великим језичким моделима, на пример за аутоматско формулисање речничких дефиниција — приступ примењен на словеначке фразеолошке јединице у раду ([Gantar, 2024](#)). Сличан поступак већ је спроведен и над грађом PCCJ и укључен у базу, али засад статички; у наредним фазама планира се његово понављање са разрађенијим упитима (промптовима) и новијим моделима, уз динамичко повезивање са речничком базом. Речничка база представља поуздан извор лексичког знања који може бити коришћен као спољна база знања у RAG архитектурама.

У таквом окружењу кориснички упит се најпре прослеђује API слоју који проналази релевантне речничке чланке. Након тога се добијени садржај користи као контекст за језички модел, чиме се значајно повећава поузданост генерисаних одговора.

Поред примене у оквиру RAG архитектуре, API PCCJ може бити доступан (енгл. *exposed*) путем MCP (Model Context Protocol) сервера, чиме би речник постао директно доступан агентским платформама као што су ChatGPT, Claude и Gemini.

Развој интерфејса и корисничких апликација представља природан корак у трансформацији речника из традиционалног лексикографског производа у дигиталну језичку инфраструктуру. На овај начин речнички подаци постају доступни истраживачима, програмерима, образовним институцијама и широј јавности.

Отварање дела речничке базе кроз API омогућава њену интеграцију у образовне платформе, системе за проверу правописа, претраживаче, чет-ботове и друге језичке технологије. Истовремено, веб и мобилна апликација омогућавају корисницима да једноставно приступају речнику.

5. Закључак

Речник савременог српског језика (РССЈ) представља савремену платформу за развој и дистрибуцију лексикографског знања, засновану на корпусним подацима, стандардизованом моделу података усаглашеном са DMLex и отвореном приступу речничкој грађи. Поред централне речничке базе и лексикографског окружења, нове компоненте система — REST API, веб-апликација и мобилна апликација — значајно проширују могућности употребе речника, чинећи га подједнако доступним људима и машинама.

У раду је посебна пажња посвећена интеракцији речника и корпуса, где је предложена нотација са три врсте заграда омогућила прецизнији и аутоматизован превод речничких израза у корпусне упите. Тиме је показано како се традиционална лексикографска пракса може формализовати и уградити у дигитални систем, а да се притом сачува богатство изворне речничке грађе.

Будући развој обухватиће проширење функционалности API-ја, увођење напредних механизма претраге, развој iOS верзије мобилне апликације, као и интеграцију са другим језичким ресурсима и системима вештачке интелигенције — било кроз RAG архитектуру, било преко MCP сервера. На тај начин РССЈ има потенцијал да прерасте у централну дигиталну лексикографску инфраструктуру за српски језик, способну да подржи истраживачки рад, образовање, развој језичких технологија и ширу јавну употребу.

Захвалница

Пројекат се реализује уз подршку иницијативе „С љубављу“ (<https://sljubavlju.com>), коју је покренула и координира српска дијаспора. Посебну захвалност дугујемо добротворном удружењу „Окупљени око језика“ за иницирање израде описног речника српског језика и континуирану подршку том подухвату. Развој језичких модела и технологија коришћених у пројекту подржао је Фонд за науку Републике Србије у оквиру пројекта ТЕСЛА (#7276, „Text Embeddings – Serbian Language Applications“).

Литература

- Apolonija Gantar. 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 Saša Marjanović, editor, *Moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi*, volume 1 of *Leksikografski susreti*, chapter 12, pages 151–157. Univerzitet u Beogradu, Filološki fakultet, Beograd. 12.
- Adam Kilgarrieff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. [The Sketch Engine: Ten Years on](#). *Lexicography*, 1:7–36.
- Adam Kilgarrieff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý. 2008. [GDEX: Automatically Finding Good Dictionary Examples in a Corpus](#). In *Proceedings of the XIII EURALEX international congress*, volume 1, pages 425–432. Universitat Pompeu Fabra Barcelona.
- Adam Kilgarrieff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. [The Sketch Engine](#). In *Proceedings of the XI EURALEX International Congress*, pages 105–116. Universite de Bretagne-Sud, Faculte des lettres et des sciences humaines.
- Annette Klosa-Kückelhaus and Carole Tiberius. 2024. [The Lexicographic Process Revisited](#). *International Journal of Lexicography*, 38(1):1–12.
- Iztok Kosem, Simon Krek, and Polona Gantar. 2020. [Defining collocation for Slovenian lexical resources](#). *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 8(2):1–27.
- Iztok Kosem, Simon Krek, Polona Gantar, Špela Arhar Holdt, and Jaka Čibej. 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*, page 514. ELex 2021.
- Simon Krek, Primož Ponikvar, Andraž Repar, Iztok Kosem, and David Lindemann. 2025. [DMLEX on Wikibase: Legacy Dictionaries as Collaboratively Editable Dataset](#). In *Proceedings of the eLex 2025 conference*, pages 175–189.

- Cvetana Krstev. 2008. *Processing of Serbian – Automata, Text and Electronic Dictionaries*. Faculty of Philology, Belgrade.
- Cvetana Krstev, Gordana Pavlović-Lažetić, and I Obradović. 2004. [Using Textual and Lexical Resources in Developing Serbian WordNet](#). *Romanian Journal of Information Science and Technology*, 7(1-2):147–161.
- Biljana Lazić and Mihailo Škorić. 2019. [From DELA-based Dictionary to Leximirka Lexical Database](#). *INFOtheca: Journal of Information and Library Science*, 19:81–98.
- Aleksandra Marković and Ranka Stanković. 2025. [So Close but Still Far: Case Study on Application of LLMs in Idioms Identification, Definition, and Generation of Illustrative Examples](#). In *I. Kosem, M. Jakubíček, M. Medved', K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), Electronic lexicography in the 21st Century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 Conference*, pages 79–94.
- Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 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*, pages 48–56.
- Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018b. [Resource-based WordNet Augmentation and Enrichment](#). In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018)*, pages 104–114.
- Ranka Stanković, Rada Stijović, Mihailo Škorić, and Cvetana Krstev. 2025. [The Dictionary of Contemporary Serbian Language \(RSSJ\): Advanced Automation and Other Challenges](#). In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference*, Bled, Slovenia. M33.
- Carole Tiberius, Jelena Kallas, Svetla Koeva, Margit Langemets, and Iztok Kosem. 2024. [A Lexicographic Practice Map of Europe](#). *International Journal of Lexicography*, 37(1):1–28.
- Duško Vitas and Cvetana Krstev. 2012. [Processing of Corpora of Serbian Using Electronic Dictionaries](#). *Prace Filologiczne*, LXIII:279–292.

- Duško Vitas and Cvetana Krstev. 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):105–116.
- Tadej Škvorc and Marko Robnik-Šikonja. 2025. [Solving word-sense disambiguation and word-sense induction with dictionary examples.](#)
- Душко Витас, Ранка Станковић, and Цветана Крстев. 2025. [О фамилији корпуса савременог српског језика СрпКор.](#) In *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig : Thematic Collection of Papers*, pages 3–27. University of Belgrade — Faculty of Philology.
- Рада Стијовић, Ранка Станковић, and Михаило Шкорић. 2024. [Речник савременог српског језика: РССЈ.](#) In *ЈУДИГ – Књига резимеа*, page 33, Београд, Србија. Друштво за језичке ресурсе и технологије ЈеРТех. Апстракт представљен на конференцији ЈУДИГ, 21–23. новембар 2024.

The Dictionary of Contemporary Serbian Language: From a Digital Lexicographic Resource to an Open Language Infrastructure

Ranka Stanković, Mihailo Škorić, Rada Stijović

Summary

The Dictionary of Contemporary Serbian Language (RSSJ) is a lexicographic initiative aimed at developing a modern Serbian dictionary based on corpus data, contemporary lexicographic principles, and language technologies. Unlike traditional dictionaries, the RSSJ is conceived as an integrated digital infrastructure that combines a lexical database, a lexicographic working environment, corpus resources, application programming interfaces, and web and mobile applications. This paper presents the overall system architecture and the underlying data model, aligned with the DMLex standard and implemented in PostgreSQL, as well as its integration with the SrpKor corpora and the NoSketch Engine platform. Special attention is given to the concept of a REST API that enables access to lexical data for both human users and software applications, thereby connecting the dictionary with other language resources such as the Leximirka lexical database and the Serbian WordNet. The paper also addresses the web and mobile applications, authentication and access-control mechanisms, and possibilities for integration with AI systems. It demonstrates how traditional lexicography can be transformed into a modern digital infrastructure capable of supporting research, education, industrial applications, and the development of language technologies for Serbian.

Keywords: digital lexicography, lexical databases, API, web applications, mobile applications, language technologies, Serbian language.

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

<https://creativecommons.org/licenses/by/4.0/>

