https://doi.org/10.18485/lexicog_meet.2024.1.ch23
811.163.41`374:004.9
811.163.41`322

Примљен: 30. 09. 2024.
Дорађен: 06. 12. 2024.
Прихваћен: 10. 12. 2024.

Повезивање лексема морфолошких речника коришћењем базе *Лексимирка*

Биљана Рујевић[1] — **Ранка Станковић**[2] — **Михаило Шкорић**[3]
[1] https:// orcid.org/0000-0002-9103-3902, [2] https://orcid.org/0000-0001-5123-6273,
[3] https:// orcid.org/0000-0003-4811-8692

Универзитет у Београду – Рударско-геолошки факултет (Србија)

[1] biljana.rujevic@rgf.bg.ac.rs (✉), [2] ranka.stankovic@rgf.bg.ac.rs, [3] mihailo.skoric@rgf.bg.ac.rs

Сажетак

Рад приказује приступ успостављању повезивања лексема у *Морфолошким речницима српскоī језика*. Повезивање, тј. успостављање релација не би било могуће без претходне конверзије речника из облика текстуалних датотека у облик лексичке базе података назване *Лексимирка*. Методологија за успостављање релација почива на 69 појединачних релација заснованих на 388 правила. Правила за повезивање се дефинишу на основу обележја лексичких записа (врсте речи, маркера, граматичких категорија и подниски). Успостављене релације су крајњем кориснику видљиве путем апликације *Лексимирка* у форми хипервеза и могу се сврстати у три врсте: варијационе, деривационе и изговорне релације. Варијационим релацијама су повезане лексеме које међусобно представљају варијантне облике (нпр. *кафа* и *кава*, *евро* и *еуро*). У деривационе релације спадају везе међу лексемама које су повезане деривационим правилима (нпр. *колач* и *колачић*, *дугме* и *дугменце*). Изговорном релацијом су повезани облици речи екавског и ијекавског изговора (нпр. *бијел* и *бео*, *сњешко* и *снешко*). Остварено је укупно 103.589 повезивања лексичких записа (кроз 43 варијационе релације остварено 3.401 повезивање, кроз 25 деривационих релација 94.732 повезивања и једну изговорну релацију 5.456 повезивања).

Кључне речи: морфолошки речници, повезивање лексема, лексичка база података, српски језик.

Abstract

The paper presents an approach to establishing relations between lexemes in *Serbian Morphological Dictionaries* (SMD). These relations would not be possible without the prior conversion of the SMD dictionaries from text file formats into a lexical database — *Leximirka*. The methodology consists of 69 relations defined by 388 different rules that are based on lexical entry properties such as part of speech, markers, grammatical features, and substrings. The established relations are visible to the end user through the application of the same name (*Leximirka*) in the form of hyperlinks and can be categorized into three types: variational, derivational, and pronunciation relations. Variational relations connect lexemes representing variant forms of each other, such as *kafa* and *kava* (coffee), *euro* and *evro* (euro), etc. Derivational relations include connections between lexemes linked by derivational rules (e.g., *kolač* (cake) and *kolačić* (cookie), *dugme* (button) and *dugmence* (little button)). Pronunciation relations connect word forms used in the ekavian and ijekavian pronunciations, such as *bijel* and *beo* (white) or *snješko* and *sneško* (snowman). Using these relations, 103,589 lexical entries pairs are established (3,401 using 43 variational relations, 94,732 using 25 derivational relations and 5,456 using 1 pronunciation relation).

Keywords: morphological dictionaries, lexeme connection, lexical database, Serbian language.

1. Увод

Морфолошки речници српскоī језика (Krstev 2008) представљају за српски језик битан електронски језички ресурс за обраду природних језика који се развија више од три деценије. Подаци који се бележе у лексичким записима чине драгоцен извор различитих лексичких информација. Лексички записи првобитно су чувани у текстуелном формату, након чега је извршена трансформација у лексикографску базу података названу *Лексимирка*[1] (Lazić – Škorić 2020). Ова база омогућила је обогаћивање речника подацима који нису били саставни део формата, попут података о фреквенцијама о коришћењу речи у корпусима, као и низ функционалности које су обогатиле речник и проширили му намену (Рујевић 2022).

Овај рад приказује развој једне од нових функционалности речника, која се односи на развој система за повезивање лексема кроз варијационе, деривационе и изговорне релације. Подаци представљени

[1] Истоимена апликација је доступна на адреси: <https://leximirka.jerteh.rs> уз логовање с налогом електронске поште регистрованим на платформи Гугл. Корисници су у могућности да претражују *Морфолошке речнике српскоī језика* и прегледају податке о лексичким записима.

у раду односе се на најажурнији пресек стања овог система заснован на 69 различитих релација и 388 правила уз помоћ којих је успостављено 103.589 повезивања лексичких записа (варијационе релације: 3.401 повезивање; деривационе релације: 94.732 повезивања; изговорна релација: 5.456 повезивања). Сама методологија успостављања релација ослања се на постојање лексичке базе података и најчешће аутоматско остваривање повезивања. Повезивање је засновано на дефинисању правила која почивају на ознакама лексичких записа углавном специфичним за *Морфолошке речнике српскоī језика*. Ознаке које се користе за креирање правила чине врста речи, деривациони, релациони и семантички маркери, те граматичке категорије. Аутори сматрају да је овај, махом аутоматски систем заснован на правилима, ефикасан за обогаћивање *Морфолошких речника српскоī језика* информацијом која претходно није постојала, а од значаја је за кориснике. У појединачним проблематичним случајевима, описаним у одељку 5, долази до ручног повезивања лексичких записа или раскидања веза. У пракси се, кроз развијену апликацију за преглед и управљање речником, успостављено повезивање манифестује постојањем хипервезе међу лексичким записима међу којима је остварена веза.

Везе о којима је реч у овом раду постоје и у традиционалним речницима српског језика. Релација која повезује екавски и ијекавски изговор у РМС и РСАНУ представљена је на тај начин што се најчешће у оквиру речничког чланка екавског изговорног лика наводи и ијекавски (скраћеницама *ијек.* и *јек*.). У ретким случајевима ситуација је обрнута. У предговору првом тому РСАНУ наведено је да су речи јужног изговора унете у РСАНУ као засебне одреднице али без дефиниција, примера и других података (само се у првом тому користи скраћеница *ј*). Са њих се упућује на исте речи екавског изговора, где се ијекавски ликови обрађују као тип дублета. У речничким чланцима екавског лика дају се сви подаци с дефиницијама и примерима. Ради уштеде простора, ијекавски ликови који припадају гнезду, тј. већој породици речи с истим „кореном" не уносе се као одреднице, већ се као одредница наводи „корен" с цртицом, који упућује на исти екавски корен. Из истог разлога се у РСЈ екавски ликови праћени (и)јекавским изговором уводе скраћеницом *јек*.

У РМС, као и у РСЈ, варијантни облици речи с гласовним разликама или различитим префиксима или суфиксима, представљају се тако што се дефиниција даје у речничком чланку уобичајеног облика док остали облици упућују ка том облику. Други начин за представљање ових облика огледа се кроз упоредне одреднице у оквиру истог речничког чланка. У РСАНУ ови типови дублета повезују се тако што се код стандардног облика даје дефиниција с примерима, док се код других облика дају примери њихове употребе. За упућивање на друге облике користи се скраћеница *в*. (види).

Деривациони облици у сва три поменута традиционална речника повезују се преко дефиниције тако што се остали облици дефинишу на основу основног облика. Тако је у РМС уз реч радник дата пуна дефиниција док је облик *радница* дефинисан као „жена радник", а реч *раднички* као „који се односи на раднике, који је у вези са радницима".

С обзиром на то што су ови речници још увек у папирној форми, међу речничким чланцима не постоји хипертекстуалност.

Структура је овога рада таква да је у одељку 2 дат преглед информација о *Морфолошким речницима српскоī језика*. С посебном пажњом је описан изворни формат њихових лексичких записа, битан за разумевање развијеног система за повезивање. Одељак 3 описује лексичку базу података на примеру дела предвиђеног за успостављање веза међу лексичким записима. Одељак 4 описује методологију за успостављање релација међу лексичким записима и то према појединачним врстама релација. У посебним пододељцима описана је методологија за успостављање варијационих (4.1), деривационих (4.2) и изговорне релације (4.3). У одељку 5 дат је приказ неких проблема уочених приликом успостављања релација. Одељком 6 илустрована је практична употреба система за управљање релацијама и њихов приказ кроз апликацију *Лексимирка*. У одељку 7 дата су нека закључна разматрања у погледу значаја развијених повезивања за *Морфолошке речнике српскоī језика*.

2. Морфолошки речници српског језика

Морфолошки речници српскоī језика (Krstev 2008) представљају електронске речнике намењене, пре свега, употреби у рачунарским апликацијама које се баве обрадом природних језика. Они су значајан језички ресурс за језике с богатом флексијом, у које се сврстава и српски језик. Почивају на формату DELA (*Dictionnaires électroniques du LADL*), који је развијен у лабораторији LADL (*Laboratoire d'Automatique Documentaire et Linguistique*) под руководством М. Гроса (Maurice Gross). Развој и примена *Морфолошких речника за српски језик* почињу пре тридесетак година и везани су за научноистраживачки рад Д. Витаса (Vitas 1993) и Ц. Крстев (Krstev 1997), који и данас руководе развојем и допуном речника. Систем *Морфолошких речника за српски језик* почива на теорији коначних аутомата (Vitas 2006), па је заснован

на морфолошким и локалним граматикама представљеним коначним трансдукторима (енг. *Finite State Transducer, FST*), којима се дефинишу и генеришу сви флективни облици речи у речницима.

Морфолошки речници DELA састоје се од речника монолексемских јединица (енг. *simple words*) и речника полилексемских јединица (сложених речи — енг. *compounds*, или вишечланих јединица одн. израза — енг. *multiword units*, *MWU* или *multiword expressions*, *MWE*).

Основне компоненте морфолошких речника монолексемских речи чине DELAS (фр. *DELA de formes simples*) и DELAF (фр. *DELA de formes Fléchies*) (Krstev 1997). Компонента DELAS састоји се од леме одговарајуће речи и флективног, семантичког и синтаксичког обележја. Следи пример краћег записа из речника монолексемских јединица:

(1) mineral,N1+Conc+Mat+DOM=Geol

Облик *mineral* представља лему, потом иза запете следе ознака флективне класе *N1*. Семантички маркер *Conc* означава да се ради о конкретној именици, маркер *Mat* означава да се ради о материјалу, док доменски маркер *DOM = Geol* означава да термин припада домену геологије.

Леме са придруженим флективним кодом, тј. флективним трансдуктором из речника DELAS, омогућавају да се аутоматски генеришу одреднице у речнику DELAF. Записи речника DELAF састоје се од облика речи, леме, ознаке врсте речи и граматичких категорија. Следи опис једног флективног облика *mineralom* претходно наведене леме *mineral*.

(2) mineralom,mineral.N +Conc+Mat+DOM=Geol:ms6q

Иза флективног облика и леме наведена је врста речи која је у овом случају именица означена кодом *N*. Потом следе маркери речника DELAS, који су исти као и у претходном примеру. Иза двотачке следе ознаке граматичке категорије *m* за мушки род, *s* за једнину као ознаку броја, *6* као ознака за падеж инструментал и *q* као ознака аниматности која показује да реч има неаниматно (неживо) својство. Речник DELAF генерише се аутоматски из речника DELAS и флективних аутомата.

Основне компоненте система морфолошких речника полилексемских речи чине DELAC (фр. *DELA de formes composés*) и DELACF (фр. *DELA de formes composées fléchies*). Следи запис за полилексемску јединицу *rudni mineral* из речника DELAC:

(3) rudni(rudni.A2:adms1g) mineral(mineral.N1:ms1q),NC_AXN+DOM=Geol

Ова лема састоји се од две компоненте *rudni* и *mineral*. У пратећим заградама су дати описи флективног облика који формирају ову полилексемску јединицу у датом канонском облику (номинатив једине). Ово значи да је облик *rudni* део флективне парадигме леме придева *rudni* који припада флективној класи *A2* и представља позитив (*a*), одређеног придевског вида (*d*), мушког рода (*m*), једнине (*s*), у номинативу (*1*) и без обележја аниматности (*g*). Компонента *mineral* представља именицу чија је лема у истом облику и мења се у складу са флективном парадигмом *N1*. Облик је у мушком роду (*m*) једнине (*s*) номинатива (*1*) без обележја аниматности (*q*). Иза леме праћене знаком за запету следи ознака флективне класе полилексемског израза *NC_AXN*. Ова флективна класа означава да се полилексемска реч састоји од именице којој претходи придев који се слаже са именицом у роду, броју, падежу и обележју аниматности, док се граматички број полилексемске речи не мења, односно остаје исти као у леми. Потом следи доменски маркер *DOM=Geol* који показује да реч припада домену геологије.

И најзад, следи запис из речника DELACF који описује један флективни облик *рудне минерале* полилексемске јединице *рудни минерал* која је претходно представљена у речнику DELAC:

(4) rudne minerale,rudni(rudni.A2:adms1g) mineral(mineral.N1:ms1q).NC:mp4q

Ради се о облику полилексемске именице (*NC*) који је мушког рода (*m*) у множини (*p*), акузативу (*4*) и неаниматан (*q*). Записи речника DELACF се генеришу аутоматски уз помоћ речника DELAC и флективних аутомата монолексемских и полилексемских јединица. Вишезначни флективни облици полилексемских јединица дефинишу се једним редом у речнику DELACF са више група граматичких категорија које га дефинишу.

Примена је *Морфолошких речника срūскоī језика* вишеструка, почев од основних задатака обраде текста коришћењем система *Unitex*[2], кроз постављање различитих сложених упита регуларним изразима или графовима (коначним аутоматима) како би се из текста екстраховали различити подаци или да би се обавила нека сложена трансформација текста. Поред тога, речници се примењују и при различитим задацима од који су неки аутоматско препознавање термина у различитим доменима, препознавање именованих ентитета, препознавање временских израза, екстракција података из доменских текстова, претрага података у дигиталним библиотекама, корекције текста враћањем дијакритичких знакова итд.

Према подацима из августа 2024. године, речник монолексемских јединица састоји се од преко 240.000

[2] Више о систему: https://unitexgramlab.org/.

лема, док је број полилексемских јединица обухваћених речником 22.875. Најзаступљеније врсте речи чине именице (116.192 лема), придеви (64.274 лема) и глаголи (21.159 лема).

Одржавање морфолошких речника првобитно је спровођено кроз подсистем радне станице за управљање лексичким ресурсима *WS4LR* (*Work Station for Lexical Resources*) (Stanković 2009), која је касније прерасла у софтверски алат *Лексимир*. Могућности које је *Лексимир* пружао (Stanković et al. 2011) биле су дистрибуција речника у више датотека, претраживање и издвајање подскупова лема на основу различитих критеријума који су саставни део DELA формата. Даље, омогућавао је везу с коначним трансдукторима и регуларним изразима који описују флексију дате леме, што је корисно из два разлога. Први је прегледање и кориговање флективних трансдуктора, уколико за тиме има потребе. Други је могућност генерисања свих облика нове леме, што је значајно за проверу одабира кода флективне класе. *Лексимир* је имао и специјалне могућности за рад с речницима у формату DELAC, а пре свега формирање речника DELAC на основу листе полилексемских јединица, што је омогућено моделом који предвиђа исправну флективну класу полилексемске јединице, као и облике њених компонената (Krstev – Vitas 2009).

Имајући у виду да је број лексичких записа нарастао, као и број особа које раде на развоју *Речникâ*, јавила се потреба за апликацијом која ће подржати истовремени вишекориснички рад, што алат *Лексимир* није могао да пружи. Таква апликација морала је бити заснована на лексичкој бази података о којој ће бити реч.

3. Модел базе података и њено формирање

Како би само повезивање лексема, односно лексичких записа *Морфолошких речника срūскоī језика* било могуће било је неопходно податке из речника из текстуелног формата претворити, односно похранити у базу података. Како би се дошло до базе података која би задовољила лексикографске потребе, односно омогућила каснију поновну употребу података за различите потребе, размотрена су три стандардизована модела представљања лексичких података: Иницијатива за обележавање текста (енг. *Text Encoding Initiative*), Оквир за лексичко обележавање LMF (енг. *Lexical Markup Framework*) и модел *lemon,* развијен као стандард за дељење лексичких информација на семантичком вебу. Иницијатива за обележавање текста је највише намењена означавању података из традиционалних речника иако у оквиру заједнице која га користи постоји велико интересовање за повезивање обележених података са Отвореним повезаним подацима (енг. *Linked Open Data*) (Bański et al. 2017), те она није узета као модел. Имајући у виду природу *Морфолошких речника* одлучено је да се база података моделује имајући у виду комбинацију модела LMF и *lemon*. Треба имати у виду да је модел *lemon* најмлађи и да је развијен са узором на модел LMF, који представља међународни стандард, али уз коришћење мање елемената и краћих назива класа и њихових својстава (McCrae et al. 2012). Уз ово *lemon* модел тежи коришћењу екстерних лексикона за различите потребе, нпр. за опис морфологије, као и онтолошком приступу у дефинисању значења. Посебно је значајно што се овај модел динамично развија. Током 2019. године су се појавили модули за лексикографију (енг. *The OntoLex Lemon Lexicography Module — lexicog*) и модул за информације о фреквенцијама, потврдама и корпусима (енг. *Module for frequency, attestation and corpus information — FrAC*) (Chiarcos et al. 2020). Модел за лексикографију настао је из потребе да се постојећи лексички ресурси објављују у виду отворених повезаних података (Bosque-Gil et al. 2019), што и јесте главни правац у развоју савремене лексикографије.

Приказ 1 илуструје упрошћени извод дела модела лексичке базе података који се користи за успостављање релација међу лексичким записима. У табели *LexicalEntry* похрањује се највећи део лексичких података из формата речника *DELAS*. С обзиром на то да се приказани модел лексичке базе односи на успостављање релација, у овој табели налазе се подаци о појединачним лексичким записима међу којима је успостављена релација. Овде су смештене информације о леми, врсти речи и флективној класи.

(5) mineralizacija,N600+DOM=Geol+VN+Process

На *ūприказу 1* дат је пример на записима *минерал* и *минерализација*, од којих је први приказан као пример (1) у претходном одељку. Дакле, подаци о лемама *минерал* и *минерализација* (пример 5) и врстама речи, тј. да су записи именице *N,* као и ознаке флективних класа *N1* и *N600* смештени су као записи у овој табели. Ова табела је посредством две везе (са два повезана лексичка записа) повезана са табелом *LexicalRelation*, која је потом повезана с табелом *DCValueRelRules*, у којој се налазе појединачна правила за успостављање релација. Правило које треба да задовоље записи с приказа дефинише да полазна и циљана реч у повезивању треба да буду именице (*POS=N*) и да крајња реч у повезивању садржи афикс *изацūја*. Правила дефинисана у табели *DCValueRelRules* иначе представљају критеријуме на основу врсте, флективне класе или маркера којима треба да буде означен лексички запис за повези-

вање. Како би релација била недвосмислено дефинисана, потребни су подаци складиштени у табели *DatCatValuesRelations*. У овој табели се похрањују кодни подаци који недвосмислено идентификују категорије података које се користе у речнику. На приказу видимо да је категорија за врсту речи именица идентификована као *id=1, ord=1*, док је семантички маркер који означава процес идентификован као *id=10058, ord=44*. Више правила за повезивање из *DCValueRelRules* може бити везано за једно правило из табеле *DatCatValuesRelations*.

Табела *LexicalSense* представља везу лексичког записа и онтологије, а у случају морфолошких речника чини везу лексичког записа и његових маркера синтаксичко-семантичких значења. Један лексички запис може имати више значења, те је на приказу веза табеле лексичког записа с табелом лексичког значења исказана као један према више. Код примера (5) постоји једно значење *+DOM=Geol+VN+Process*. Сваки појединачни маркер похрањује се у табели *SenseProperties*. Дакле, појединачни маркери за домен, глаголску именицу и процес наћи ће се у овој табели. Табела *SenseRelation* предвиђена је за повезивање појединачних значења лексичких записа са синонимима и ворднетом.

4. Методологија повезивања лексема

Повезивање лексема извршено је на основу унапред дефинисаних правила заснованих на обележјима лексичких записа из *Морфолошких речника*. За повезивање су коришћене ознаке врста речи, семантички и деривациони маркери, као и подниске садржане у лемама лексичких записа. У зависности од тога који се критеријуми користе за повезивање, саме релације могу се сврстати у три групе, и то варијационе релације (пододељак 4.1), деривационе релације (пододељак 4.2) и изговорне релације (пододељак 4.3). Стога је детаљна методологија илустрована примерима дата у наредним пододељцима.

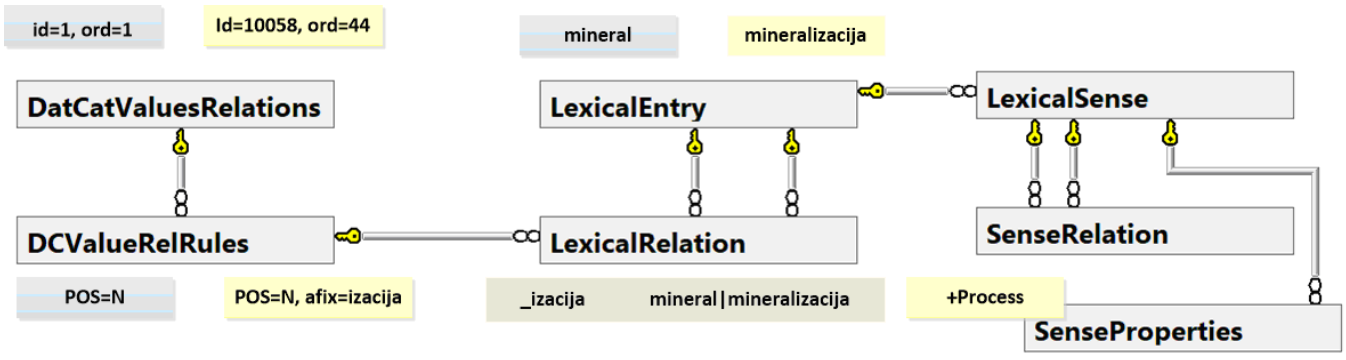
4.1. Варијационе релације

У варијационе релације спадају релације којима се повезују лексички записи који чине језичке варијанте речи истог значења нпр. *едуковати* и *едуцирати*, *кафа* и *кава*, *цертификат* и *сертификат*. Успостављене релације углавном се односе на дијатопијску варијацију. Аутори нису узимали у обзир финију разраду, имајући у виду да је повезивање вођено деривационим маркерима који су присутни у *Морфолошким речницима српског језика*. Стога предмет рада нису биле релације које би повезале парове лексема попут *прозор* и *пенџер* или дијафазне релације које би повезале пар *кева* и *мајка*. Да би се успоставила варијациона релација, потребно је да лексички записи претходно буду обележени одговарајућим маркерима деривације.

Следе лексички записи за леме *едуцирати* (пример 6) и *едуковати* (пример 7) повезани варијантном везом *cirati_kovati*, као и лексички записи за леме *кафа* (пример 8) и *кава* (пример 9) повезани варијантном везом f_v:

(6) educirati,V1+Imperf+Perf+Tr+Iref+DER= KovatiCirati
(7) edukovati,V18+Imperf+Perf+Tr+Iref+DER= CiratiKovati

При успостављању релације између два глагола *едуцирати* и *едуковати*, правилом је дефинисано да један лексички запис (пример 6) садржи маркер *+DER=KovatiCirati*, што значи да садржи подниску *цирати*, која у другом лексичком запису може бити замењена подниском *ковати*. Други лексички запис за повезивање (пример 7) означен је маркером *+DER=CiratiKovati*, што значи да подниска *ковати* може бити замењена подниском *цирати*. Правилом је предвиђено да обе речи буду глаголи, тј. да врста речи буде обележена ознаком *V*. Овим правилом означеним као varijanta(*cirati_kovati)* повезано је 13 парова лексема, међу којима су *електрифицирати* и *електрификовати*, *индуцирати* и *индуковати* или *фабрицирати* и *фабриковати*.



Приказ 1. Модел дела лексичке базе *Лексимирка* који се односи на релације.

(8) kafa,N600+DOM=Culinary+DER=FV+Conc+Drink+Food+Prod
(9) kava,N600+DOM=Culinary+DER=VF+Conc+Drink+Food+Prod

Лексички запис за лему *кафа* (пример 8) је обележен варијационим маркером *+DER=FV* који означава да подниска *ф* може бити замењена подниском *в*, док је други запис леме *кава* (пример 9) означен маркером варијација *+DER=VF* који представља знак да подниска *в* може бити замењена подниском *ф*. Врста речи није од значаја за ову релацију, за разлику од претходног примера. На основу овог правила успоставља се релација именована као varijanta(f_v). Овим правилом су успостављене 53 релације међу лексичким записима, а примери повезаних парова су: *кефтати* и *кевтати*, *салфета* и *салвета*, *кафотија* и *кавотија*, *куглоф* и *куглов*.

Извод неких од варијационих релација дат је у *табели 1*. У првој колони дат је назив релације, у другој и трећој колони приказани су маркери којима треба да буду означени полазни и крајњи лексички запис за повезивање. У четвртој колони наводи се број појединачних правила која дају кандидате за повезивање, док је у петој колони дата продуктивност релације у виду броја успостављених релација међу лексичким записима. У последњој колони представљени су примери успостављених повезивања на основу критеријума релације из датог реда. У првом реду табеле представљена је варијациона релација која подразумева да је полазни запис означен маркером *DER=ArisatiIrati* док је крајњи запис обележен са *DER=IratiArisati*. Ова релација је описана помоћу три правила која појединачно додатно дефинишу да оба записа за повезивање треба да буду иста врста речи: код једног правила именице, другог глаголи и трећег придеви. У последњој колони видимо да су на овај начин повезани парови *коментарисање* и *коментирање*, *прокоментарисати* и *прокоментирати*, као и *коментарисан* и *коментиран*. За потребе успостављања повезивања варијационим релацијама развијено је 90 различитих правила. Број правила по релацији варира од 1 до 6, а најбројније су оне које су засноване на по једном правилу. Оне обично подразумевају тип релације који подразумева замену једног слова међу повезаним лексемама. Примери таквих релација дати су у последња три реда у табели. Међу најпродуктивнијим је релацијама варијационог типа релација, којом се остварује 678 повезивања и која је заснована на 3 правила и подразумева маркере *DER=AvatiIvati* и *DER=IvatiAvati* на лексичким записима за повезивање. Видимо да су неки од парова повезаних овом релацијом *осмишљаван* и *осмишљиван* као и *укисељавање* и *укисељивање*.

4.2. Деривационе релације

У групу деривационих релација спадају оне релације које повезују лексичке записе што су повезани на основу правила заснованих на деривационим маркерима из *Речника*. На овај начин повезане су лексеме: *радник* и *радница* (релација — моција рода), те *радник* и *раднички* (релација — релациони придев). Ове релације су карактеристичне за одређене врсте

Релација	Полазна вредност	Крајња вредност	Бр. правила	Бр. повезивања	Примери
varijanta	DER=ArisatiIrati	DER=IratiArisati	3	6	komentarisanje – komentiranje; prokomentarisati – prokomentirati; komentarisan – komentiran
varijanta	DER=AtiIrati	DER=IratiAti	3	5	izmiksan – izmiksiran; miksanje – miksiranje;
varijanta	DER=AtiOvati	DER=OvatiAti	3	34	špikan – špikovan; jodlati – jodlovati
varijanta	DER=AvatiIvati	DER=IvatiAvati	3	678	osmišljavan – osmišljivan; ukiseljavanje – ukiseljivanje
varijanta	DER=CiratiKovati	DER=KovatiCirati	3	115	educiran – edukovan; publicirati – publikovati
varijanta	DER=CS	DER=SC	1	6	certifikat – sertifikat; sufinanciran – sufinansiran
varijanta	DER=FV	DER=VF	1	91	salfeta – salveta; kuglof – kuglov
varijanta	DER=HJ	DER=JH	1	40	čoha – čoja; snahin – snajin; smeh – smej

Табела 1. Примери правила и повезивања код варијационих релација.

речи. Следећи лексички записи илуструју наведене релације:

(10) radnik,N10+Hum
(11) radnički,A2+PosQ
(12) radnica,N651+Hum+GM

Правило за успостављање релације *relacioni pridev* заснива се на повезивању именице и придева и постојању маркера *+PosQ* на придеву. Правило којим је успостављено повезивање пара *радник* (пример 10) и *раднички* (пример 11) захтева подниску *к* код именице (пример 10) и подниску *чки* код придева (пример 8). Овим правилом су повезана 52 пара речи а неке од њих су: *запӣсник* и *запӣснички*, *дневник* и *дневнички*, *речник* и *речнички*. Релација моције рода између записа *радник* (пример 10) и *радница* (пример 12) остварује се на основу критеријума да обе речи морају бити именице и да једна реч садржи маркер *+GM* (пример 12) који означава моцију рода. Прва реч треба да садржи подниску *к* (пример 10) док друга реч треба да садржи подниску *ца* (пример 12). Подниска може бити на било којој позицији у речи, иако се у овом примеру налази на крају речи. Овим правилом успостављено је 59 релација међу паровима речи од којих су неки: *срећник* и *срећница*, *вереник* и *вереница*.

У *та̄бели 2* видимо да је међу најпродуктивнијим релацијама она која повезује презиме са присвојним придевом именована као *prisv. pridev — prezime* са 17.280 повезаних парова речи. Видимо да полазна реч у релацији треба да буде обележена маркером за презиме *+Last*, док крајња реч треба да буде означена маркером за присвојни придев *+Pos*. Овим правилом су повезани парови: *Андрић* и *Андрићев*, *Тинто̄рето̄* и *Тинто̄рето̄в*. До сада је развијено 25 различитих деривационих релација, које су засноване на 278 правила. Овде је број правила по релацији у просеку доста већи него код варијационих релација и варира од 1 до 43. На по једном правилу засновано је 9 релација а њих 6 засновано је на преко 20 правила. Релација заснована на највећем броју правила, чак 43, која повезује именицу са деминутивом представљена је у претпоследњем реду табеле.

4.3. Изговорна релација

Изговорна релација између екавског и ијекавског лика речи успоставља се на основу маркера за изговор — екавски *+Ek* и ијекавски *+Ijk*, којима су обележени лексички записи. Врста речи није од значаја за ову релацију. Следи пример лексичких записа из *Морфолошких речника српскоī језика* повезаних релацијом *Ek–Ijk*:

(13) beo,A38+Col+Ek
(14) bijel,A14+Col+Ijk

Правило којим су повезани лексички записи *бео* (пример 13) и *бијел* (пример 14) дефинише да један лексички запис мора садржати подниску *ео* (пример 13) док други мора садржати подниску *ијел* (пример 14). Овим правилом је успостављено 7 релација. Неки од повезаних парова су: *цео* и *цијел* и *по̄лубео* и *по̄лубијел*.

Релација	**Полазна вредност**	**Крајња вредност**	**Бр. правила**	**Бр. Повезивања**	**Примери**
glagolska imenica	Imperf	VN	18	7391	blistati — blistanje; misliti — mišljenje
glagolska imenica	Perf	VN	10	1842	izlečiti — izlečenje; odocneti — odocnjenje;
prisvojni pridev – prezime	Last	Pos	24	17280	Andrić — Andrićev; Tintoreto — Tintoretov
prisvojni pridev – ime	First	Pos	1	51	Mijailo — Mijailov; Vujo — Vujov; Jerotije — Jerotijev
mocija roda prezime	Last	GM	2	19764	Petersen — Petersenka; Mrgodić — Mrgodićka
relacioni pridev	nema	PosQ	37	10244	Nant — nantski; violina — violinski
prisvojni pridev	nema	Pos	18	12183	prijatelj — prijateljev; kedar — kedrov
mocija roda	nema	GM	8	1370	amater — amaterka; znanac — znanica; lutkar — lutkarica
deminutiv	nema	Dem	43	1786	brod — brodić; stopalo — stopalce
augmentativ	nema	Aug	10	193	kupus — kupuščina; sin — sinčina

Табела 2. Примери правила и повезивања код деривационих релација.

(15) prepis,N1+Ek
(16) prijepis,N1+Ijk

У случају повезивања записа *препис* и *пријепис*, правило је да један лексички запис садржи подниску *е* (пример 15) док други мора да садржи подниску *ије* (пример 16). На овај начин је успостављена 101 веза а неки од повезаних парова су: *осмехнути* и *осмијехнути*, *петао* и *пијетао*, *бедник* и *биједник*. Изговорна релација заснована је на 20 различитих правила којима је начињено 5.456 повезивања.

5. Потешкоће уочене приликом повезивања лексема

Приказани систем функционише врло ефикасно имајући у виду да је могуће за кратко време успоставити велики број повезивања и да је највећи број релација, чак 103.472 од 103.589, успостављен аутоматски. Без обзира на то, уочили смо поједине ситуације које можемо окарактерисати као потешкоће приликом успостављања релација међу лексичким записима. У тим ситуацијама долази до повезивања лексема која су неисправна. Ти случајеви се у начелу могу класификовати кроз три сценарија: лексички записи нису обележени одговарајућим маркером који је предуслов за успостављање повезивања; у морфолошком речнику недостаје лексички запис за упаривање; или, услед широко постављеног правила, долази до погрешног повезивања лексема, тј. лексичких записа.

Када лексички запис није обележен одговарајућим маркером који је предуслов за успостављање повезивања, постоји могућност његове ручне допуне одговарајућим маркером чиме се решава проблем. Друга могућност која је пригодна приликом допуне више лексичких записа јесте поновно покретање аутоматског успостављања релације кроз сегмент апликације за успостављање релација. Код релације *imenica_proces* која повезује именице са одговарајућим именицама које означавају процесе на основу задовољавања критеријума да су обе лексеме именице и да друга лексема садржи суфикс *изација*, а лексички запис садржи семантички маркер *+Process,* примећено је да је овај маркер недостајао. Након допуне лексичких записа *пелетизација*, *минерализација*, *карбонатизација*, *калцитизација* овим маркером, остварена су повезивања парова: *карбонат* и *карбонатизација, калцит* и *калцитизација, минерал* и *минерализација, пелет* и *пелетизација.*

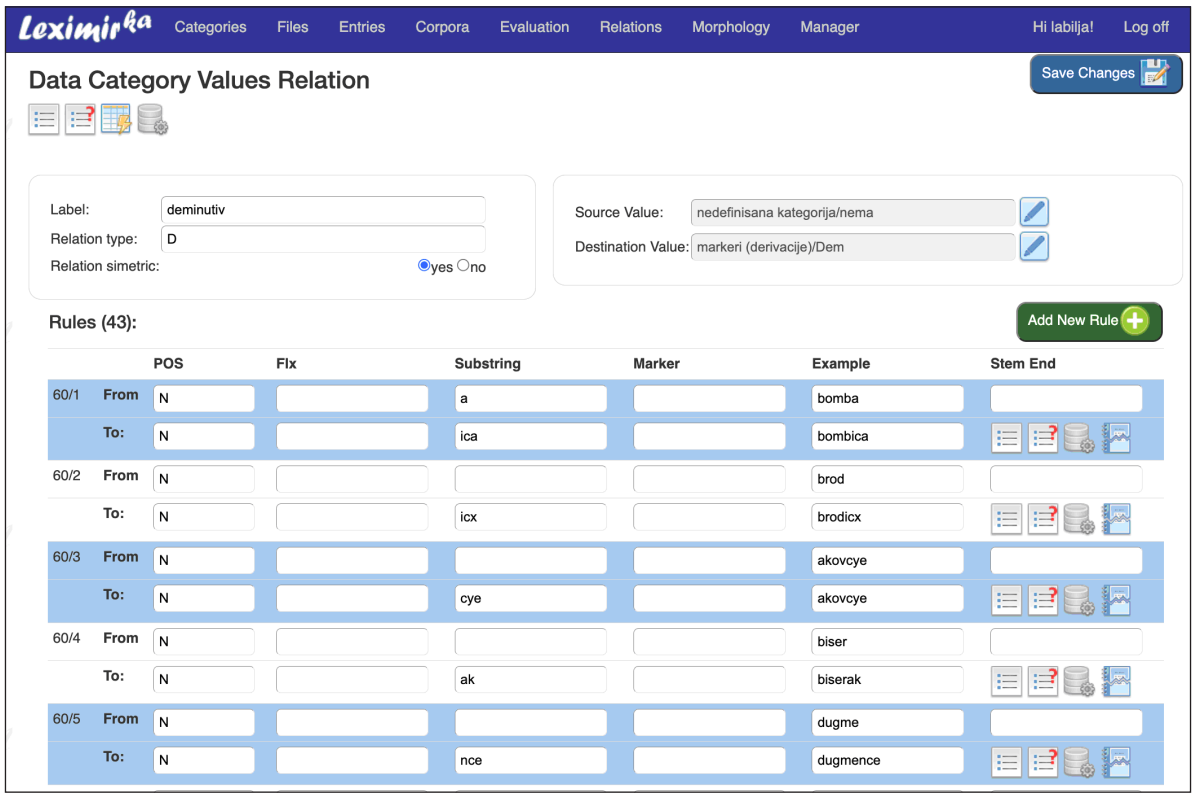
Када у *Морфолошким речницима* нема лексичког записа с којим је могуће остварити упаривање, проблем се решава простом допуном речника изосталим записом. То представља добар начин за допуну самог речника. Овај случај можемо сагледати кроз пример релације *relacioni pridev (_ski)*, која повезује релациони придев што се завршава на суфикс *ски* са именицом. У пракси су приликом повезивања лексема *афиолит* и *афиолитски*, *алб* и *албски*, недостајале лексеме које представљају именицу, те је речник допуњен записима који представљају одговарајуће именице (*афиолит* и *алб*) како би повезивање било остварено.

Пример погрешног повезивања услед широко постављеног правила илустроваћемо кроз повезивање *Вран* и *Врањанка* путем релације *ženski stanovnik*, која спаја топониме са становницима, правилом које одређује да се подниска *н* у једној речи мења подниском *њанка* у другој речи. За одређивање топонима и становника користе се семантички маркери *+Top* и *+ Inh*. Прва именица именује планину у Босни и Херцеговини док друга именује становницу града Врања. Овако погрешно успостављена повезивања ручно се раскидају. До сада је 71 релација означена као погрешна, те је као таква и ручно раскинута. Ручно раскидање релација спроводи се из интерфејса за уређивање лексичког записа на једном од два погрешно повезана записа.

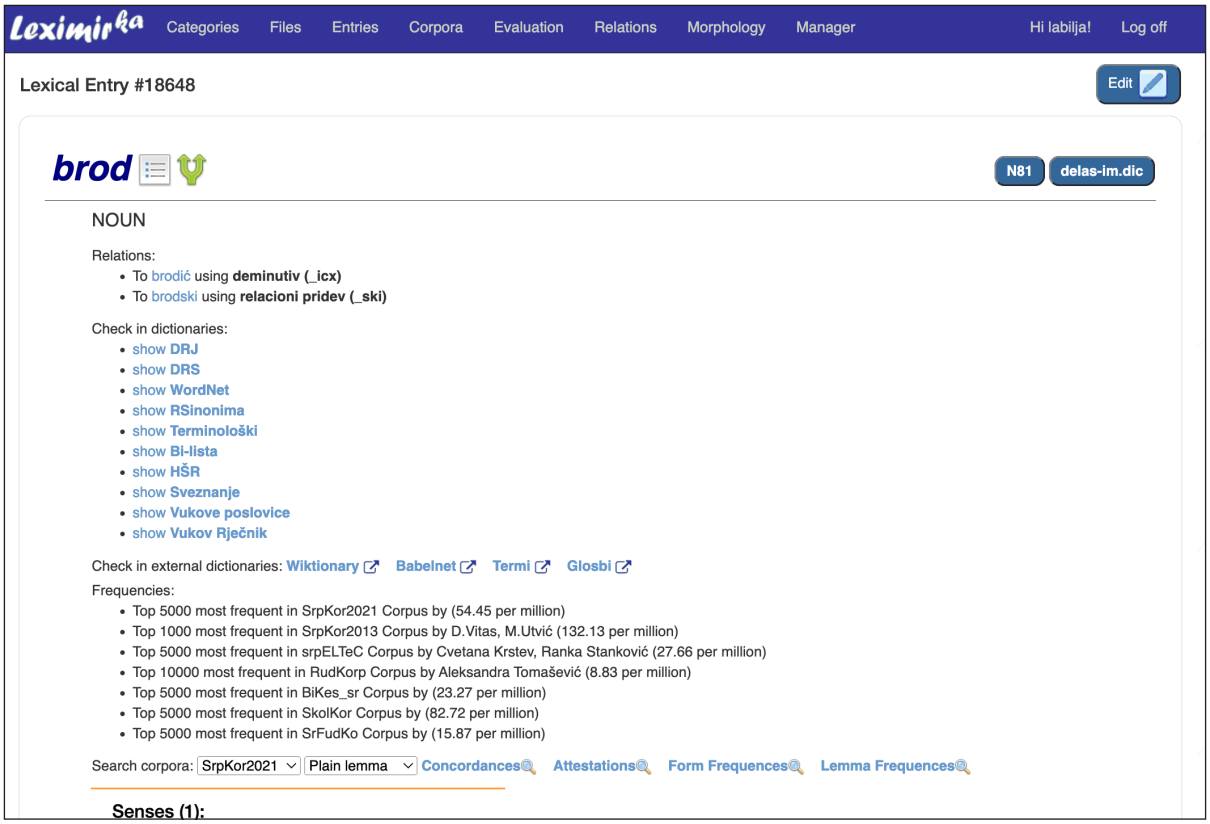
6. Приказ релација кроз апликацију *Лексимирка*

У пракси се успостављање релација спроводи кроз сегмент апликације *Лексимирка* намењен управљању лексичким релацијама назван *Relations*. *Приказ 2* даје снимак екрана за управљање релацијом *deminutiv*. Извршење свих правила релације покреће се коришћењем групе дугмади испод назива *Data Category Values Relation*. Ова дугмад се редом односе на приказ повезаних лексема, приказ нових кандидата за повезивање, приказ брзих кандидата и извршење правила за успостављање релације. Уз помоћ прва два дугмета могуће је проверити исправност извршених повезивања и ваљаност креираног правила и сходно томе проценити да ли је приступ исправан. Следећи сегмент осенчен сивом бојом односи се на административне податке о релацији, тј. саму ознаку и врсту, као и на критеријум који треба да се задовољи (на нивоу целе релације) да би се повезивање испунило. У плаво и бело осенченим редовима дата су појединачна правила која чине ову релацију. Прва колона односи се на правила везана за врсту речи, друга на флективну класу, трећа на подниску, четврта на маркер. У петој колони дати су примери парова који су повезани на основу датог правила. Дугмад дата уз појединачна правила имају врло сличне функције као она која су дата на нивоу релације, само што служе за извршење на нивоу правила. Разлика је у последњем дугмету, дугмету за брисање правила. Корисник може додати и дефинисати ново правило коришћењем дугмета *Add new rule*. На *приказу 2* видимо да је првим осенченим

редом дефинисано прво од 43 правила за успостављање релације *deminutiv*, којим је дефинисано да су полазна и крајња лексема за повезивање именице, да прва треба да садржи подниску *а*, док друга треба да садржи подниску *ица*. Као пример је наведен пар *бомба* и *бомбица*.



Приказ 2. Пример приказа панела за управљање релацијом.



Приказ 3. Пример приказа лексичког записа кроз апликацију *Лексимирка*.

На *приказу 3* налази се пример лексичког записа за лему *брод* приказан кроз апликацију *Лексимирка*. У првој групацији података излистане су везе с другим лексичким записима, као и информација о врсти релације и правилу на основу кога је она успостављена. Видимо да је успостављена веза са записом *бродић* коришћењем раније поменуте релације *deminutiv,* правила према коме полазна реч не треба да садржи подниску, док се на циљаној речи додаје суфикс *ић*. Друга успостављена релација је *relacioni pridev* и остварена је са лексемом *бродски*, правилом код кога се код циљане лексеме очекује суфикс *ски*. Поред сегмента са релацијама, ваља напоменути да се на приказу виде и друге функционалности апликације које се односе на претрагу леме у различитим речницима, информације о њеним релативним фреквенцијама у различитим корпусима, као и претрага различитих корпуса чистом лемом (*plain lemma*) или предефинисаним обрасцима који садрже лему. Резултати претраге корпуса могу се приказати у виду конкорданци, потврда, фреквенција облика леме или фреквенција саме леме.

7. Закључна разматрања

Представљена правила за повезивање лексичких записа заснована су на формату морфолошких речника и имплементирана захваљујући постојању лексичке базе података. Као критеријуми за успостављање правила за повезивање коришћени су подаци о врсти речи лексема које се повезују, семантички и деривациони маркери којима су означене, као и подниске садржане у лемама.

Предност представљене методологије за успостављање варијационих, деривационих и изговорних релација може се огледати у ефикасности ако се узме у обзир да је уз помоћ једног дефинисаног правила за повезивање просечно повезано 266 парова речи, што је податак који се добија када се подели број успостављених релација (103.589) са бројем дефинисаних правила (388). Једна од највећих предности овог приступа огледа се у томе што се успостављена правила за повезивање могу користити као база знања о релацијама међу речима српског језика. Уз то, систем је проширив, тако да је могуће додавање нових релација и допуна постојећих правила.

Ново окружење *Морфолошких речника српскоī језика Лексимирка* има за циљ да олакша употребу речника и прошири њихов круг корисника. С обзиром на то да су ознаке коришћене у лексичким записима приказане и на природном језику, да је, између осталог, извршено повезивање лексичких записа с различитим корпусима и речницима на интернету и у локалној лексичкој бази података, очекује се да речник могу користити и лексикографи као средство за рад. Успостављене релације међу лексичким записима видљиве у облику хипервеза приказане у овом раду такође обогаћују корисничко искуство.

Напомена

Истраживање спроведено уз подршку Фонда за науку Републике Србије, број гранта 7276, (Text Embeddings - Serbian Language Application – TESLA).

Литература

Рујевић, Биљана (2022). *Речници у дигиталном добу - информатичка подршка за српски језик* (необјављена докторска дисертација). Београд: Филолошки факултет.

*

Bański, Piotr, Bowers, Jack, Erjavec, Tomaž (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In: Iztok Kosem et al. (eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 485–494). Brno: Lexical Computing CZ s.r.o.

Bosque-Gil, Julia, Gracia, Jorge, McCrae, John, Cimiano, Philipp, Stolk, Sander, Khan, Fahad, Depuydt, Katrien, de Does, Jesse, Frontini, Francesca, Kernerman, Ilan (2019). *The OntoLex Lemon Lexicography Module*. <https://www.w3.org/2019/09/lexicog/#introduction>. [13. 8. 2024]

Chiarcos, Christian, Ionov, Maxim, de Does, Jesse, Depuydt, Katrien, Khan, Fahad, Stolk, Sander, Declerck, Thierry, McCrae, John (2020). Modelling Frequency and Attestations for OntoLex-Lemon. In: Ilan Kernerman et al. (eds.), *Proceedings of the 2020 Globalex Workshop on Linked Lexicography* (pp. 1–9). Marseille: European Language Resources Association.

Krstev, Cvetana (1997). *Jedan prilaz informatičkom modeliranju teksta i algoritmi njegove transformacije* (необјављена докторска дисертација). Beograd: Univerzitet u Beogradu, Matematički fakultet.

Krstev, Cvetana (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Belgrade: Faculty of Philology of the University of Belgrade.

Krstev, Cvetana, Vitas, Duško (2009). An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds. In: B. Lamiroy et al. (eds.), *The 28th Conference on Lexis and Grammar, Bergen, 29th September - 3rd October 2009*, *In Arena Romanistica* (pp. 204–212). Bergen: University of Bergen, Department of Foreign Languages.

Lazić, Biljana, Škorić, Mihailo (2020). From DELA Based Dictionary to Leximirka Lexical Database. *Infotheca — Journal for Digital Humanities*, *19*(2), 81–98.

McCrae, John, Aguado-de-Cea, Guadalupe, Buitelaar, Paul, Cimiano, Philipp, Declerck, Thierry, Gómez Pérez, Asunción, Gracia, Jorge, Hollink, Laura, Montiel-Ponsoda, Elena, Spohr, Dennis, Wunner, Tobias (2012). *The Lemon Cookbook*. < http://lemon-model.net/lemon-cookbook.pdf >. [13. 8. 2024]

Stanković, Ranka (2009). *Modeli ekspanzije upita nad tekstuelnim resursima* (необјављена докторска дисертација). Beograd: Univerzitet u Beogradu, Matematički fakultet.

Stanković, Ranka, Obradović, Ivan, Krstev, Cvetana, Vitas, Duško (2011). Production of morphological dictionaries of multi-word units using a multipurpose tool. In: K. Jassem et al. (eds.), *Proceedings of the Computational Linguistics-Applications Conference* (pp. 77–84). Warsaw: Polish Information Processing Society.

Vitas, Duško (1993). *Matematički model morfologije srpskohrvatskog jezika (imenska fleksija)* (необјављена докторска дисертација). Beograd: Univerzitet u Beogradu, Matematički fakultet.
Vitas, Duško (2006). *Prevodioci i interpretatori : (uvod u teoriju i metode kompilacije programskih jezika)*. Beograd: Matematički fakultet BU.

Лексикографски извори

РМС: Стевановић, Михаило (ур.) (1967–1976). *Речник српскохрватскога књижевног језика*, *1–6*. Нови Сад — Загреб: Матица српска, Матица хрватска.
РСЈ: Николић, Мирослав (ур.) (2011). *Речник српскога језика*. Нови Сад: Матица српска.
РСАНУ: Белић, Александар и др. (ур.) (1959–2023). *Речник српскохрватскога књижевног и народног језика. 1–22*. Београд: Институт за српскохрватски језик САНУ.

Biljana Rujević — Ranka Stanković — Mihailo Škorić

Establishing relations between lexemes of morphological dictionaries using the *Leximirka* database

(Summary)

The paper outlines a methodology for linking lexemes in Serbian Morphological Dictionaries (SMD) through a lexical database, *Leximirka*. The need for this lexical database arose from the limitations of the dictionary format based on text files. By converting these files into a structured database, the *Leximirka* platform allows for the automatic generation and management of various relations between lexical entries. These relations enhance the functionality of Serbian Morphological Dictionaries, making them more accessible and useful for humans and natural language processing (NLP) applications. The core types of relations established in *Leximirka* include variational, derivational, and pronunciation-based relations. Variational relations connect lexemes that are different forms of the same word, such as *kafa* and *kava* (coffee) or *euro* and *evro* (euro). This type of relation captures minor differences in spelling or form without changing the meaning. Derivational relations, on the other hand, connect lexemes based on shared roots and grammatical transformations, linking words like *radnik* (worker) and *radnica* (female worker), or *kolač* (cake) and *kolačić* (cookie). These relations are useful for understanding how different word forms are generated in Serbian, a language with a rich inflectional system. Pronunciation relations account for the differences between the two main pronunciations of Serbian: Ekavian and Ijekavian. Examples of this relation include *beo* and *bijel* (white), or *snješko* and *sneško* (snowman), which are pronounced differently but share the same meaning. The *Leximirka* web application makes these relationships visible to users through hyperlinks, offering an intuitive and interactive way to explore connections between lexemes. This not only improves the user experience, but also aids in various tasks in which dictionaries are used. Additionally, the web application facilitates the management and expansion of the dictionaries by supporting multi-user collaboration and relation-building based on predefined lexical rules. Overall, *Leximirka* significantly enhances the utility of Serbian Morphological Dictionaries, enabling more sophisticated lexicographic research and supporting various NLP applications. It represents a valuable tool for the ongoing development and modernisation of Serbian language resources, especially considering that it is one of the languages with less developed language resources.