

Ивана Лазић Коњик*

(Институт за српски језик САНУ, Београд)

ЛЕКСИКОГРАФИЈА У ДИГИТАЛНОЈ ЕРИ: ИНТЕРНЕТ И ДРУШТВЕНИ МЕДИЈИ КАО ИЗВОРИ НОВИХ РЕЧИ**

Новоречје, 2024, бр. 11

У раду се разматрају питања везана за концепт интернета и друштвених медија као лексикографског корпуса, узимајући у обзир хипотезу да интернет представља богат репозиторијум података који се успешно може користити у лексикографске сврхе. Описује се пракса новопокреног пројекта израде онлајн речника неологизама српског језика (Нео21) који користи интернет као корпус за прикупљање, селекцију и документовање нових речи, као и у изради речничког чланка. Циљ овог речника је да корисницима омогући брз и једноставан приступ речничким подацима, пружајући најактуелније и поуздане информације и прецизна објашњења значења. У закључку се истиче да су интернет и друштвени медији важан и актуелан извор за праћење и документовање језичких промена, као и репрезентативан корпус за израду нових типова речника који би пратили развој језика у реалном времену.

Кључне речи: интернет, друштвени медији, лексикографски корпуси, речници нових речи, Нео21, лексикографија, српски језик.

Увод

1.0 Са појавом и развојем интернета и друштвених медија,¹ пред лексикографе су постављени нови изазови. С једне стране, традиционални

* ivana.konjik@isj.sanu.ac.rs; ORCID: 0000-0003-0840-0140

** Овај рад је финансирао Министарство за науку, технолошки развој и иновације Републике Србије према уговору број 451-03-66/2024-03/ 200174 који је склопљен са Институтом за српски језик САНУ у Београду.

¹ Термини *друштвени медији* и *друштвене мреже* се у српском језику користе као синоними да означе платформе које омогућавају корисницима да активно креирају и деле различите врсте садржаја (текстове, коментаре, слике, видео снимке, линкове и др.) и да се међусобно повезују и комуницирају једни са другима (Инстаграм, Фејсбук, Јутјуб, Твитер/Икс, Тикток и др.), с тим да је у свакодневној комуникацији у српском језику за

медији, који своје садржаје објављују и на интернету, постали су лакше доступни за праћење и прикупљање језичких података, при чему поред текстова које садрже и штампана издања, уводе нов начин комуницирања са својим читаоцима путем блогова, коментара читалаца и сл. С друге стране, појава друштвених медија донела је праву револуцију у области писаног језика, стварајући нов језички варијетет који се, по својим грама-тичким и лексичким особинама, разликује како од конвенционалног писа-ног језика, тако и од говорног језика.

1.1 Слобода изражавања на друштвеним медијима, у блоговима, тви-товима, објавама, коментарима и др., где припадници широке јавности у буквалном смислу могу да напишу шта год желе, довела је до бројних је-зичких иновација, посебно лексичких. Поред великог броја нових речи ко-је свакодневно улазе у српски језик, најчешће из енглеског,² присутне су и разне друге лексичке новине, као што су скраћенице, акроними, сливени-це, емотикони итд.

1.2 Истраживања показују висок проценат употребе неологизама на друштвеним мрежама (в. рад Šetka Čilić, Ilić Plauc 2021, у којем су анали-зирани неологизми са четири друштвене мреже: Facebook, Twitter, Insta-gram и WhatsApp). Учестала појава неологизама делимично је последица чињенице да интернет и друштвени медији омогућавају брже усвајање но-вих речи од стране њихових корисника и њихово продирање у језичке идиоме ван друштвених мрежа.³ Иако већина поменутих лексичких инова-ција није забележена у речницима, многе од њих су се укорениле у свако-дневној језичкој употреби припадника ове медијске заједнице.

1.3 Прикупљање и записивање нових речи и израза са свим њиховим иновативним карактеристикама: фонетско-фонолошким, граматичким,

ово значење уобичајенији други назив, друштвене мреже. У стручном регистру, посебно у енглеском језику, међутим, прави се одређена разлика између значења ова два термина. Термин *друштвени медији* односи се на саме платформе за креирање и дељење различи-тих садржаја, док се термин *друштвена мрежа* односи на апликацију која омогућава по-везивање корисника у заједнице на овим платформама и њихову међусобну комуникаци-ју (исп. Радивојевић 2024, Википедија: друштвени медији).

² Проблем огромног и пребрзог уплива речи из енглеског језика за које немамо адекватну замену, више пута је разматран у литератури, а посебно у радовима Т. Прћића, в. нпр. Ргџић 2019, 2023.

³ Пре појаве интернета, појављивање нових речи се најбоље могло пратити на страницама штампе, јер су новине и часописи први извештавали о новим догађајима. Та-ко су ове речи брзо стизале до шире публике, чиме су се стабилизovala њихова значења у језику (Мжељскаја и Степанова 1983: 127, према Оташевић 2008: 11). Новине су и даље оптималан извор података о фреквентној употреби неологизама јер се објављују свако-дневно и тако могу брже да прате промене у језику него други писани материјали (Creese 2017: 3).

творбеним и семантичким представља веома важан припремни процес за њихово укључивање у лексички фонд српског језика и њихово распоређивање у различите његове делове (основне или периферне) и употребно-вредносне позиције (потенцијалне речи, okazaionalizmi и неологизми), а даље за утврђивање њихове стабилне употребе и укључивање у референтан и ажуриран речник савременог српског језика, који ће пружати свеобухватан увид у актуелан језички развој и његове промене.

Традиционалне методе

2. Традиционалне методе прикупљања нових речи подразумевале су ручну ексерпцију неологизама из писаних извора, штампаних публикација, новина и часописа, евентуално из телевизијских и радио емисија. Потом се вршила провера да ли су ове лексеме забележене у постојећим речницима општег типа, речницима нових речи, речницима страних речи и др. релевантним изворима (исп. Muhvić-Dimanovski 2005: 109; Оташевић 2004р: Увод; Šipka 1998: 140).⁴

Корпусна лингвистика

3.0 Развој рачунара и дигиталних технологија довео је до повећања капацитета за чување и обраду података. Педесетих и шездесетих година 20. века почиње да се развија корпусна лингвистика која снажан замах доживљава у последњим деценијама 20. века (Lenders 2013). Употреба рачунарске технологије омогућила је аутоматизовано састављање и обрађивање великих текстуалних база, а у новије време, интернет и дигитални ресурси омогућили су стварање веома великих корпуса, који садрже милијарде речи (Fuentes-Olivera 2012: 52).

3.1 Крајем 20. века, паралелно са развојем корпусне лингвистике развијају се софтверски алати за лингвистичку анализу. Детекција нових речи почела је да се ослања на статистичке анализе и поређење са постојећим речницима. Унапређивање рачунарских програма омогућило је аутоматизованију анализу и издвајање потенцијално нових речи и израза из корпу-

⁴ Поред ових уобичајених метода, у традиционалном приступу ређе су заступљене методе спровођења анкета међу различитим групама говорника како би се открила заступљеност нових речи које се користе у одређеним социјалним групама или професијама. Разговори са појединцима који су у додиру са специфичним жаргоном или техничким језиком, такође могу бити драгоцени за идентификацију нових термина. Специјализовани извори, као што су научни радови и техничка документација, могу помоћи у проналажењу нових термина, нпр. у разним техничким и научним областима. Нови закони и прописи такође доносе нову политичку и друштвену терминологију која улази у стандардни језик. Научни радови, мастерске и докторске тезе и др., који се баве неолошким питањима, добар су извор нових речи (Lazić Konjik, Milenković 2020: 481).

са. Захваљујући развоју машинског учења и великих језичких модела развијени су софтверски алати који препознају нове лексеме на основу контекста и фреквенцијских анализа (као што је Sketch Engine⁵). Ови модели сада могу да идентификују нове речи и неологизме на основу њихове употребе у различитим контекстима на друштвеним мрежама, форумима, блогovima и другим онлајн изворима, омогућавајући лексикографима да брже прате и бележе језичке промене (нпр. NeoCrawler, алат који аутоматски прати и прикупља неологизме са интернета, Kerremans и др. 2102; исп. и Krek 2024). На тај начин се у лексикографски процес израде дигиталних речника осим лексикографа укључује низ других стручњака, као што су, корпусни лингвисти, рачунарски лингвисти, програмери, дизајнери, успостављају се стандарди и формати кодирања речника, као што су TEI Lex-0, ONTOLEX-LEMON, ISO LMF (Lexical Markup Framework) ISO 1951:2007, DMLEX (Klosa-Kückelhaus, Tiberius 2024: 65).

3.2 Друштвене мреже, као што су Јутјуб, Фејсбук, Твитер и Инстаграм, посебно су корисне за уочавање неологизама, јер омогућавају директан увид у језик који користи широка јавност. Корисници који креирају садржаје на овим платформама спонтано прихватају и стабилизују употребу језичких иновација кроз различите активности, репостовање, хештегове и друге сличне комуникативне праксе, што омогућава истраживачима да лакше идентификују актуелне језичке трендове. Употреба неологизама на платформама друштвених медија обично претходи њиховом ширењу у другим, традиционалним изворима и на вебу (Würschinger 2021), док њихово ширење у друге домene употребе може бити важан показатељ у процесу њихове конвенционализације (Würschinger i др. 2016: 43).

Стање у српском језику

4. Иако постоје алати и методе који би се вероватно могли прилагодити за откривање нових речи у српском језику,⁶ главни проблеми са којима се суочавају српски лексикографи у примени дигиталних алата и метода за праћење нових речи и ажурирање речника огледају се у неколико кључних фактора. Прво, то је недостатак ажурних корпуса. Постојећи корпуси српског језика не ажурирају се, због чега не одражавају у потпуности савремени језички развој, посебно када је реч о неологизмима који настају

⁵ <https://www.sketchengine.eu/>

⁶ Језичким технологијама за српски језик бави се група истраживача окупљена око ЈеРТеха, Друштва за језичке ресурсе и технологије, в. <https://jerteh.rs/> и <https://jerteh.rs/index.php/resursi-i-alati/>

у домену друштвених мрежа и нових медија.⁷ Друго, не постоји сарадња између лексикографа и информатичара. Лексикографи нису обучени за рад са сложеним дигиталним алатима и технологијама које омогућавају екстракцију и анализу нових речи из великих текстуалних података. Овај недостатак сарадње успорава развој алата прилагођених специфичним потребама лексикографа и доводи до тога да се доступни алати недовољно користе. Процес идентификације нових речи по правилу се ослања искључиво на ручне методе које обављају лексикографи, које су споре и мање ефикасне. Треће, развој дигиталних корпуса, као и обука лексикографа за рад са новим алатима, захтевају стабилна улагања у инфраструктуру и људске ресурсе. Међутим, лексикографија у Србији нема довољно финансијске подршке да би могла пратити технолошке иновације на пољу обраде језика. Све су то разлози због којих се у српском језику у изради речника и даље прибегава ручним методама у изради корпуса и праћењу појаве нових речи, што показује кратка критичка анализа методологије прикупљања нових речи рецентних речника нови(ји)х речи и англицизама, која следи.

Примери из праксе: Оташевић 2004, РНА, СРНА, РНРуСЈ, НЕО21

5.0 Досадашња пракса указује на недостатке у инфраструктури, сарадњи са информатичарима и технолошкој подршци. Примери наведених рецентнијих српских речника нови(ји)х речи и англицизама показују да је

⁷ Постоји неколико корпуса српског језика, најстарији је *Корпус савременог српског језика* Математичког факултета (2003) који обухвата приближно 23 милиона речи и (2013) са приближно 122 милиона речи (уз регистрацију доступни на <http://www.korpus.matf.bg.ac.rs/korpus/mainmenu.php>). У новије време појавили су се слободно доступни веб корпуси: srWas (2016) настао повлачењем текстова са интернета 2014. са приближно 477 милиона речи (<https://www.clarin.si/repository/xmlui/handle/11356/1063>); PDRS (Serbian Web) (2023) Института за српски језик САНУ са приближно 603 милиона речи (<http://hdl.handle.net/11356/1752>), настао повлачењем текстова са интернета 2022. год.; Serbian web corpus CLASSLA-web.sr 1.0 (2024) са приближно 3 милијарде 343 милиона речи (<http://hdl.handle.net/11356/1931>), настао повлачењем текстова са интернета 2021. и 2022. године. Такође постоји и корпус ЈерТеха (Друштва за језичке ресурсе и технологије) са преко 800 милиона речи који није јавно доступан. Ови корпуси генерално су релевантни за праћење савремених лингвистичких промена. Најстарији корпуси, попут корпуса савременог српског језика Математичког факултета, мање су корисни за праћење (нај)новијих неологизама, јер се не ажурирају. Међутим, новији веб корпуси, попут PDRS и CLASSLA-web.sr, који представљају пресек актуелног стања за време у коме су настали, покривају актуелније изворе и боље одражавају трендове, па су релевантнији за анализу нових речи. Пошто се ни они не ажурирају, њихова значајност за праћење најновијих речи се смањује. Без могућности аутоматске екстракције нових речи, рад са таквим корпусима је мање ефикасан, будући да лексикографи морају „ручно“ тражити нове речи.

методологија прикупљања и анализе нових речи заснована превасходно на традиционалним методама и ручном раду лексикографа.⁸

5.1 *Велики речник нових и незабележених речи* Ђ. Оташевића (2004) (штампани речник) показује да је методологија прикупљања грађе за овај речник нових речи обухватила комбинацију ручног рада и аутоматизованих техника. Један део грађе сакупљан је у периоду од 1984. до 1997. године и заснивао се на делимичном ексцерпирању текстова из штампаних извора, часописа, дневних листова и књига. Овај ручни метод прикупљања подразумевао је пажљиво одабирање и издвајање нових речи из читаних материјала, што је изискивало велики труд и време. С друге стране, коришћене су и аутоматске методе у изради лексикографског корпуса. Коришћењем компјутерских програма, са интернета су преузети текстови који су претворени у облик погодан за лексикографску обраду. Ова техника омогућила је ефикасније и брже прикупљање и документовање нових речи, у поређењу са ручним сакупљањем. Укључена је и дигитализација кроз скенирање целих књига значајних српских аутора и научника, чиме је проширена грађа доступна за анализу. Иако су примењене различите методе, примарни фокус остаје на ручном раду при сакупљању и одабиру коначних одредница за речник, упркос употреби аутоматизованих алата за прикупљање текста са интернета.

5.2 RNA (2001) (штампани речник) је такође заснован на традиционалним методама и садржи искључиво материјал из штампаних медија (дневних и недељних новина). То значи да се ослањао на ручно ексцерпирање, што је утицало на величину и разноликост корпуса. У то време, интернет је био у раној фази развоја, због чега није постојала могућност да се прате нове речи које настају на интернету или друштвеним мрежама.

5.3 SRNA (2021) (дигитализовани штампани речник)⁹ је, с друге стране, користила дигиталне изворе: материјал онлајн медија и текстове са интернета, са информативних, образовних, забавних и других портала и вебсајтова, као и објаве на друштвеним мрежама, форумима и у блоговима на српском језику. Међутим, ни у изради SRNE није постојала сарадња са

⁸ За неке речнике приметан је недостатак методолошких информација (нпр. RNA), због чега није сасвим јасно како су прикупљани подаци. Изузетак представља речник англицизама SRNA, за који је у објављеном приручнику *Практични лексикографски водич* јасно и детаљно наведена методологија рада.

⁹ Под дигитализованим штампаним речником подразумевамо речник који је направљен као класична књига и који је објављен у ПДФ формату на интернету и доступан за преузимање. Такав речник не поседује интерактивне или динамичне функције, као што су претраживање и редовно ажурирање, па се не може сматрати правим онлајн речником, иако је у електронском формату. Овакви речници задржавају статичну форму (оригиналног) штампаног издања, иако су доступни корисницима путем интернета.

информатичарима. Прикупљање и обрада језичког материјала рађени су ручно, радом лексикографа (Прћић 2023: 31). Сарадници, који су за то били нарочито обучени, претраживали су интернет и прегледали (ишчитава-ли) и ексцерпирали примере са англицизмима. Т. Прћић наводи да је у процесу стварања корпуса за СРНА, у одабиру, испису и обради корпу-сног материјала, учествовало 10 обучених сарадника који су ексцерпцију вршили према писаном упутству о поступку ексцерпције материјала (Пр-ћић 2023: 32).¹⁰ За сам процес прикупљања лексема, илустративан је њи-хов коментар који се односи на методологију рада: „Поред тога, поједини сарадници правили су пописе англицизама с којима би се сусретали у ра-зним ситуацијама у току дана – тако би се, најчешће у мобилни телефон, уносили лексеме које би се чуле у разговору с пријатељима, у јавном пре-возу, на телевизији и сл., или пак оне које би се уочиле на билбордима, натписима, рекламама, па чак и на етикетама производа. Када би се село за рачунар, лако би се, употребом Гугл-претраживача, пописани англици-зми пронашли у контексту (у реченици, односно делу реченице), а потом унели у корпус, чиме би се допунио списак параграфа добијених регулар-ним путем – ишчитавањем текстова на интернету“ (Шафер и др. 2023: 86).

5.4 Методологија континуираног праћења и бележења нових речи и њиховог накнадног проверавања на интернету доследно је примењена и у РНРуСЈ (2022, 2023) (дигитализовани штампани речник). Аутори нису, ка-ко напомињу у Предговору речника, кренули од општег корпуса, него су саставили списак одредница из различитих извора: из новина, са телевизи-је, радија, друштвених мрежа, из књижевности, научне (неолошке) и науч-нопопуларне литературе, административних текстова, за које су затим тра-жили примере употребе на интернету. Ова методологија је добро осми-шљена, али се и она у великој мери ослања на ручни рад, због чега је при-мењени приступ веома захтеван у смислу времена и ресурса. Такође може довести до пропуштања нових речи које нису одмах препознате или које се појављују у мање уобичајеним контекстима. С друге стране, дигитали-зовани штампани формат речника ограничава могућност благовременог објављивања нових речи, јер захтева дуг процес припреме и објављивања, што често резултира кашњењем у одражавању актуелних језичких проме-на.

5.5 Слична методологија прикупљања нових лексема примењује се и у оквиру новопокренутог пројекта израде првог *онлајн речника* неологиза-

¹⁰ Процес ексцерпирања грађе за SRNA детаљно је представљен у раду *Прикупљање, одабир и обрада грађе за Српски речник новијих англицизама* (Шафер и др. 2023). У тексту се прецизно наводе извори који су кориштени, обухваћене тематске области и ме-тодологија рада.

ма, који ће обухватати период 21. века, Нео21 (2024).¹¹ Непостојање ажурног корпуса, с једне стране, и нагли, свакодневни прилив нових речи, с друге, наметнули су потребу редовног праћења и прикупљања „најсвежих“ речи које још нису ушле ни у корпусе ни у речнике српског језика. Четворочлана група сарадника на овом пројекту „лови“ нове речи које се појављују на интернету и друштвеним медијима, а затим их објашњавају и документују одговарајућим материјалом са интернета и на основу доступних рецентних корпуса српског језика, узимајући у обзир ширину распрострањености и фреквенцију употребе. Сарадња са информатичарима огледа се у развоју софтвера за онлајн речник. Ипак, ни у овом пројекту није било могуће коришћење дигиталног корпуса и аутоматског издвајања нових речи из претходно наведених разлога, због чега се прикупљање и даље заснива на ручном раду и сталном мониторингу дигиталних извора. Вреди нагласити да је извршена значајна методолошка измена у процесу прикупљања и одабира нових речи. За разлику од традиционалног лексикографског приступа који је захтевао припрему корпуса, овај приступ избегава ту фазу, чиме се значајно убрзава процес рада, а као примарни извор користи се интернет¹².

5.6 Анализа претходних речника, поред тога што је показала да се методологија прикупљања нових речи у српском језику и даље углавном ослања на ручни рад, истовремено је указала на значај интернета и друштвених медија као кључних извора у савременом контексту за праћење језичких иновација и документовање најновијих лексичких појава у српском језику.

¹¹ Грађа за овај речник објављује се у виду месечних прилога на сајту Филолошког друштва „Речи“ у рубрици Нове речи <https://www.reci.org.rs/2024/02/22/rubrika-nove- reci/>, а у најскорије време појавиће се онлајн речник под називом Српски речник неологизама - 21. века, за који је планирано да буде редовно ажуриран. У данашње време, како констатује Т. Прћић, речнике неологизама вреди реализовати само као онлајн издања због могућности брзог ажурирања у складу са савременим темпом живота. Периодично уношење нових одредница постаје неопходно због брзог појављивања нових речи (Prčić 2004: 251).

¹² Иако је корпус од кључног значаја за лексикографију, постоје речници који га не користе, попут Викиречника и других краудсорсинг речника. Ови пројекти се ослањају на доприносе заједнице уместо на систематски сакупљене језичке податке, што је покренуло расправе међу истраживачима о томе да ли би корпус могао постати више препрека него предност у дигиталној ери. У овом контексту развијају се нове ручне методологије за прикупљање података са веба, које настоје да превазиђу ограничења традиционалног коришћења корпуса и прилагоде се специфичностима савременог дигиталног окружења (Creese 2017: 7–8).

Интернет као корпус

7.0 У савременој лексикографији подаци се више не прикупљају формирањем лексикографских корпуса у виду збирке ексцерпираних листића, намењене искључиво за израду речника, као што је то било у случају штампаних речника. Уместо тога, у лексикографском раду, све више се користе општи дигитални корпуси или веб корпуси, који омогућавају приступ много ширем и разноврснијем фонду језичких података и делимичну или потпуну аутоматизацију послова. Ови корпуси се редовно ажурирају и обогаћују новим изворима, укључујући интернетске текстове и садржаје са друштвених медија, што омогућава лексикографима да, уз помоћ посебних софтверских алата за подршку лексикографским процесима, лакше прате брзе језичке промене и прикупљају актуелне примере употребе речи. Оваква пракса значајно унапређује процес израде речника, јер омогућава бржи и прецизнији увид у развој савременог језика и интеграцију нових речи и израза који одражавају актуелне тенденције у језичкој пракси.

7.1 Посебно место у формирању савремених корпуса има интернет.¹³ Све чешће се израђују тзв. веб корпуси који су базирани на интернетским садржајима. Они обухватају текстове који су аутоматски екстраховани са интернета, укључујући блогове, форуме, друштвене мреже, новинске чланке и друге онлајн изворе. Стога пружају широк спектар језичких података из различитих контекста и регистара, омогућавајући приступ различитим стиловима и типовима језичке употребе. Такође омогућавају напредне лингвистичке анализе, укључујући претрагу, фреквенцијске анализе, колокације и друге методе лингвистичке анализе (исп. Benko 2017).

7.2 Проблем са веб корпусима, као и са свим другим корпусима, јесте што морају редовно да се ажурирају. Ако се не ажурирају редовно, не могу да прате динамичне језичке промене, посебно када је у питању појава нових речи. Идеално би било да се ажурирање врши на сваких 3 до 6 месеци, али се то у пракси ретко дешава. На пример, најновији српски веб корпус PDRS 1.0 након објављивања 2022. године није обнављан новим подацима, па стога не бележи најновије речи које су се од тада појавиле и ушле у употребу.

7.3 Због ограничености статичних корпуса, у језицима у којима не постоје редовно ажурирани дигитални корпуси, лексикографи се ослањају на друге методе како би пратили и описивали језичке промене. У недо-

¹³ Однос између лексикографије и интернета је двојак. Интернет није само добар извор лексикографских података него пружа и боље основе за израду и одржавање речника (Fuertes-Olivera 2012: 52).

статку ажурираних корпуса, интернет и друштвене мреже постају основни и непосредни извор информација, служећи као својеврсни „динамички“ корпус који омогућава праћење актуелних језичких иновација у реалном времену.

НЕО21

8.0 Нови српски онлајн речник неологизама (Нео21), на чијој изради ради четворочлани тим лексикографа из Института за српски језик САНУ,¹⁴ представља пионирски подухват у модернизацији српске лексикографије, на жалост без институционалног финансирања. Овај пројекат је покренут као израз потребе за осавремењавањем српске неолошке лексикографије и прелазак на онлајн објављивање. Иако финансијска средства за рад нису обезбеђена, повод за отпочињање пројекта био је укључивање српског тима у међународну COST акцију ENEOLI, посвећену лексичким иновацијама,¹⁵ као и ентузијазам и посвећеност лексикографског тима који је уверен у његов дугорочни значај за развој лексикографије и праћење језичких промена.

8.1 Четворочлани тим лексикографа активно ради на праћењу нових лексичких појава и прикупљању материјала за израду онлајн речника неологизама 21. века. Пошто за сада не постоји могућност ослањања на аутоматизовану обраду података из дигиталних корпуса или интернета, осмишљена је нова методологија прикупљања података која комбинује позитивне резултате досадашње српске неолошке и неографске школе. Лексикографи **ручно** прикупљају податке из различитих **извора на вебу** путем Гуглове претраге: дигиталних медија, веб презентација и друштвених мрежа. Претражујемо цео веб како бисмо добили што више релевантних резултата.

8.2 Треба нагласити да природа лексике која се обрађује, као и усмереност речника ка корисницима, подразумевају тежњу да се значења нових речи или нова значења постојећих речи објасне на јасан и приступачан начин. Циљ је да их корисници не само разумеју, већ и да их могу правилно употребити у комуникацији. Овај захтев подразумева знање и ангажо-

¹⁴ Чланови пројектног тима су лексикографи из Института за српски језик САНУ, сарадници на изради великог академијског *Речника српскохрватског књижевног и народног језика* САНУ, у различитим његовим фазама израде, од основне обраде речничког текста до редакције, др Вања Миљковић, др Слободан Новокмет, др Милица Божић Синчук и др Ивана Лазић Коњик, која је иницијатор и руководилац овог пројекта.

¹⁵ У пројекту ENEOLI (<https://eneoli.eu/>) учествује истраживачки тим из Института за српски језик САНУ који чине др Ивана Лазић Коњик (МС), др Вања Миљковић и др Милица Божић Синчук.

вање лексикографа, јер машине још увек нису у стању да препознају суптилне нијансе значења, нити да их екстрахују на довољно прецизан начин.

8.3 Други кључни циљ у изради новог речника је елиминисање тзв. ефекта информационог преоптерећења, познатијег као „Google ефекат“, што такође није могуће постићи аутоматизованим приступом. Овај ефекат се јавља када корисници, претражујући одређене појмове на интернету, добију превише података, што обично доводи до потешкоћа у проналажењу тачног одговора или одустајања од претраге. Улога лексикографа је да из обиља расположивих информација сажме само релевантно, пружајући корисницима речника јасан, концизан и поуздан извор знања, у складу са лексикографским принципом „мање је више“ (Fuertes-Olivera 2012: 60). На тај начин речници постају функционални алати који задовољавају корисничке потребе.

Ограничења и предности интернета као извора

9.0 Ограничења, на која је указано у литератури, да је интернет с једне стране сувише велик и непрегледан, а са друге непоуздан извор због променљивог садржаја, што може довести до тога да примери касније не буду проверљиви (Nesselhauf, Biever 2007: 2–3, према Fuertes-Olivera 2012: 53; Lüdeling, Evert and Baroni, 2007: 11, према Creese 2017: 60), превазилазе се у нашем раду систематичним навођењем изабраног целовитог примера уз прецизне податке о извору, укључујући линк и датум објављивања. Овакав приступ осигурава да, чак и ако пример постане непроверљив због промене садржаја на интернету, наведени подаци остају релевантно сведочанство, а сам пример се у том случају може упоредити са моделским примерима који се наводе у неким речницима. Овакав методолошки поступак омогућава нам да премостимо ограничења интернета као извора, истовремено задржавајући висок ниво поузданости и научне строгости.

9.1 Интернет представља неформално и неструктурисано окружење у којем подаци нису организовани ни прилагођени за истраживачке потребе, што захтева знатан напор у филтрирању и организовању прикупљених података. Многи текстови на интернету садрже неквалитетне, ирелевантне или дуплиране податке (нпр. пренос агенцијских вести или истих текстова), што захтева додатну пажњу приликом селекције података. Различити нивои језичке правилности, од формалних до веома неформалних израза, такође утичу на квалитет прикупљених података и постављање критеријума избора за укључивање у речник (нпр. словне, правописне, граматичке и др. грешке, нестандардна ортографија, неупотребљавање дијакритичких знакова, емоције изражене типографијом, понављањем знакова, употреба

великих слова, емотикона и др., тзв. *language code switching*, замена језичког кода, најчешће српски/енглески у дискусијама, исп. и Benko 2017).

9.2 Иако претраживачи омогућавају релативно лако претраживање појединачних речи у потрази за примерима употребе, оно може бити веома отежано, нарочито када се реч поклапа са популарним именом или термином, што отежава проналажење релевантних примера. Још један изазов представља ефемерност многих израза на интернету који брзо нестају из употребе, због чега је потребан опрез приликом процене њихове трајне вредности. Поред тога, постављају се и етичка питања у вези са приватношћу и ауторским правима коришћења података с друштвених медија и других интернет платформи.¹⁶

9.3 Ипак, наша пракса је показала да интернет као извор пружа значајне предности у праћењу и бележењу језичких иновација, иако је временски и организационо захтеван и не може се у потпуности поредити са систематским радом на дигиталним корпусима.¹⁷ Његова главна предност је брзина идентификовања нових речи и израза у реалном времену и економичност (Würschinger i dr. 2016: 35). Нове језичке појаве и тенденције (нпр. адаптација позајмљеница) могу се уочити раније. Показало се да друштвене мреже посебно представљају брз и поуздан извор нових речи и неформалних израза, одражавајући актуелне језичке промене. Интернет омогућава лак и брз приступ широком спектру текстова различитих стилова и регистара језика, од формалних чланака и блогова до неформалних коментара и мемова, што пружа лексикографима детаљнији увид у савремену језичку употребу. Такође, омогућава прецизно датирање језичких иновација, што је од великог значаја за истраживање динамике неологизама.¹⁸ Интернет омогућава приступ не само језичким већ и енциклопедијским и прагматичким информацијама које су корисне за боље разумевање значења и употребе нових речи, као и стално ажурирање информација, чи-

¹⁶ Етичка питања у вези са приватношћу и ауторским правима приликом коришћења садржаја са друштвених медија су дискутована у европским оквирима, где су већ развијени одређени регулаторни механизми и смернице (в. нпр. Директиву о ауторским правима: https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_21_2821/QANDA_21_2821_EN.pdf, приступљено 22.11.2024.). Међутим, у Србији ова питања још увек нису системски регулисана, што оставља простор за будућа истраживања и формулисање националних политика.

¹⁷ Истраживачи указују да је најбоља солуција обједињавање (највећег) традиционалног дигиталног корпуса са веб корпусом за одговарајући језик (в. Benko 2017).

¹⁸ Треба напоменути да поуздане временске ознаке за веб документе нису увек доступне. У бројним случајевима постоји прецизан податак о постављању веб документа у виду целовитог датуме, док се у другим случајевима наводи само релативна ознака, нпр, пре годину дана. Поузданост гуглових временских ознака такође се може довести у питање (исп. Würschinger i dr. 2016: 37).

ме се обезбеђује релевантност и актуелност. Коначно, интернет као извор пружа обиман материјал за анализу у виду постова, твитова, блогова и других типова текстова, омогућавајући на тај начин богат корпус за израду речника који одражава динамику развоја савременог језика.

У процени степена устаљености неологизама и дефинисању њиховог значења користимо и квантитативне и квалитативне методе, чиме се осигурава свеобухватан приступ анализи.

Структура извора

10.0 У настојању да документујемо употребу нових речи, примарно се ослањамо на примере из медијског дискурса, укључујући сајтове новинских, телевизијских и радио кућа, кад год су такви примери доступни. Поред тога, користимо и садржаје са различитих веб сајтова, као што су корпоративне презентације, онлајн продавнице, странице државне управе, магацини, часописи, као и блогови, форуми, друштвене мреже и коментари корисника (в. табелу 1). Ова разноликост извора, уз примену дефинисаних критеријума селекције који обезбеђују поузданост и релевантност корпуса нових речи, омогућава свеобухватнију слику о употреби нових речи у различитим комуникационим контекстима.

10.1 Посебну пажњу посвећујемо квалитету и веродостојности извора. Искључујемо примере са сајтова са машински преведеним или генерисаним садржајем, као и дуплиране текстове који се појављују на различитим платформама. Такође, избегавамо употребу садржаја са сумњивих сајтова, попут спам и кликбејт страница,¹⁹ јер често обилују неконвенционалним језичким конструкцијама и произвољно преведеним или аутоматски генерисаним речима које немају стабилну језичку употребу и нису погодни за укључивање у речник. Треба нагласити и то да у процени неологизама водимо рачуна о посебним типовима употребе (нпр. у наслову, поднаслову, у коментару), док употреба наводника, курзива или дефинисања речи у тексту по правилу сугерише њену неолошку природу. На тај начин

¹⁹ *Спам странице* су сајтови чији је примарни циљ масовно привлачење корисника ради остваривања финансијске добити путем реклама, а не пружање квалитетних или релевантних информација. Ови сајтови садрже некавалитетан, а понекад и аутоматски генерисан садржај који је дизајниран да садржи што више популарних кључних речи и тако привуче претраживаче. *Кликбејт странице* су веб странице које користе сензационалистичке, заводљиве или манипулативне наслове и слике како би привукле кликове. Циљ је да се корисници намаме да кликну на линк због привлачног наслова, али сам садржај не пружа дубљу или релевантну информацију. Ови извори се у лексикографском истраживању углавном избегавају, јер не пружају поуздан и објективан корпус примера за документовање језика.

одржавамо висок стандард селекције, осигуравајући да корпус речника представља аутентичне примере нових речи у савременом српском језику.

Табела 1: Структура извора за Нео21

<i>врста</i>	<i>Сајтови</i>
информативни медији	ТВ: <i>admin.rts.rs; b92.net; idjtv.nova.rs; k1info.rs; nova.rs; n1info.rs; pescanik.net; pink.rs; prva.rs; rts.rs; vasseljenska.net;</i> НОВИНЕ: <i>alo.rs; bbc.com; blic.rs; danas.rs; dnevnik.rs; espreso.co.rs; euronews.rs; informer.rs; mondo.rs; nedeljnik.rs; nin.rs; novimagazin.rs; novosti.rs; nsuzivo.rs; objektiv.rs; pecat.co.rs; srpskainfo.com; standard.rs; tanjug.rs; telegraf.rs; vreme.com; zoomue.rs; 24sedam.rs;</i>
забавни медији (магазини, часописи, водичи, савети и сл.):	<i>magazin.politika.rs; elle.rs; lepotaizdravlje.rs; zadovoljna.nova.rs; pcpres.rs; buro247.rs; noizz.rs; wannabemagazine.com; wanted.mondo.rs; gloria.rs; bizlife.rs; citymagazine.danas.rs; besnopile.rs; bebac.com; penzin.rs; zelenaucionica.com; amplitudemagazin.com; najzena.alo.rs; stil.kurir.rs; tegla.rs; beforeafter.rs; dvadesete.rs; web-mind.rs; galaksijanova.rs; muskimagazin.rs; velikeprice.com; forbes.n1info.rs; stvarukusa.mondo.rs; recepti.zena.blic.rs; nadlanu.com; kockicezivota.rs; luftika.rs; injournal.rs; avlija.org.rs; eleven11eleven.rs; biznis.rs; superzena.b92.net; maxbetsport.rs; make-up.rs; demetra.rs; ekleinika.telegraf.rs; tarzanija.com</i>
личне презентације, блогови, рецепти	<i>istokpavlovic.com; ninipanini.rs; oljka.rs; bonapeti.rs; palachinkablog.com; svetskiputnik.rs; marinastaste.com; samsvojkunar.com</i>
државна управа	<i>kutak.suk.gov.rs;</i>

пословне презентације (фирми, удружења и сл.)	<i>amss.org.rs; klipin.rs; rajac.rs; protekal.rs; play.google.com; tvojvideo.rs; dijalog.net; pekariposlasticar.rs; preradamesakodmarka.rs; smartlife.mondo.rs; frizerskisonstil.rs</i>
ресторани	<i>gastrobarpark.rs; fontanapub.rs; dobridim.rs; cafebarrestoran.rs; brodzeppelin.rs</i>
продавнице	<i>dm.rs; alexandar-cosmetics.com; vasiljev.rs; mojekrpice.rs; rs.iqos.com; vapeshop.rs; decathlon.rs; zastitnaoprema.rs; impulst.net; slatkadomacica.com; delfi.rs; urbanshop.rs; bershka.com; laguna.rs; alexandar-cosmetics.com; zara.com/rs/sr/ halooglas.com; kupujemprodajem.com</i>
огласи	
школство, наука, култура, књижевност	<i>kaobalerina.rs; doi.fil.bg.ac.rs; eckermann.org.rs/; klima101.rs; reff.f.bg.ac.rs; raf.edu.rs; cpn.edu.rs; scindeks.ceon.rs; en.wikipedia.org; academia.edu; kulturkokoska.rs</i>
остале презентације	<i>izletijada.rs; rtanjskivrt.rs; rex.fondb92.org; planamedia.rs; gradsubotica.co.rs; energetskiportal.rs; vojvodina.travel</i>
друштвене мреже	<i>linkedin.com; instagram.com; x.com; facebook.com</i>
форуми	<i>ana.rs; e-cigserbia.com</i>
корпуси	<i>PDRS 1.0</i>

Закључак

11.0 Наше досадашње искуство у изради речника неологизама показало је да интернет и друштвени медији могу бити значајан извор за лексикографско истраживање нових речи у српском језику. Иако се ослањамо на ручно праћење и обраду, у недостатку аутоматизованих система, успели смо да постигнемо задовољавајуће резултате. Стекли смо одређени увид у предности које нуди интернет као корпус у односу на традиционалне лексикографске корпусе, пре свега у погледу брзине приступа информацијама и разноврсности језичких примера. Међутим, искуство је исто-

времено указало на бројне изазове, као што су филтрирање некавалитетних или нерелевантних података или решавање етичких питања у вези са приватношћу и ауторским правима приликом коришћења садржаја са друштвених медија.

11.1 С обзиром на све већи утицај дигиталних медија на језик, овај приступ несумњиво представља важан и актуелан метод за праћење и документовање језичких промена. Даљи кораци нашег рада требало би да обухвате развој или прилагођавање алата за претраживање и селекцију језичких података за српски језик, као и аутоматизацију одређених корака у обради неологизама укључујући вештачку интелигенцију. То би значајно олакшало и убрзало процес, доприносећи поузданости и ажурности речника. Главни изазов је, међутим, како постићи овај ниво интеграције и развоја у садашњим условима, што захтева не само технолошку подршку већ и институционално финансирање, као и сарадњу између лингвиста, информатичара и других стручњака. Стога апелујемо за подршку лексикографским пројектима, јер су инвестиције у лексикографске пројекте инвестиције у будућност језика.

ЛИТЕРАТУРА

- Википедија, страна Друштвени медији [https://sr.wikipedia.org/sr-ec/Друштвени_медији] приступљено 10.11.2024.
- Мжельская О. С., Степанова Е. И. 1983. Новейшие англицизмы в русском языке. У: Котелова Н. З. (ур.): *Новые слова и словари новых слов*. Москва. Стр. 125–139.
- Оташевић, Ђ. 2008. *Нове речи и значења у савременом стандардном српском језику. Лингвистички аспект*. Београд: Алма.
- Оташевић, Ђ. 2004. *Велики речник нових и забележених речи I*. Београд: Алма, 2004.
- Прћић, Т. 2023. Српски речник новијих англицизама: теоријско-методолошки и практични аспекти израде. У: Твртко Прћић, Јасмина Дражић (ур.), *Српски речник новијих англицизама од идеје до реализације. Практични лексикографски водич*. Филозофски факултет, 11–79.
- Радивојевић, С. 2024. Друштвени медији. *Етнолошко-антрополошки појмовник српске културе*. Ур. Б. Бановић. [<https://pojmovnik.rs/>] Приступљено 22.11.2024.
- РНРусЈ: Оташевић Ђ., М. Николић, В. Ђорђевић. *Речник нових речу у српском језику 1, 2*. Београд: Институт за српски језик САНУ, 2022, 2023.
- Шафер, М., А. Кардош Стојановић, М. Савић, М. Обрадовић 2023. Прикупљање, одабир и обрада грађе за Српски речник новијих англицизама. У: Т. Прћић,

- J. Дражић (ур.), *Српски речник новијих англицизама од идеје до реализације. Практични лексикографски водич*. Филозофски факултет, 81–100.
- Benko, V. 2017. Language Code Switching In Web Corpora. In: RASLAN 2017: *Recent Advances in Slavonic Natural Language Processing*. Ed. Aleš Horák, Pavel Rychlý, Adam Rambousek. Brno: Tribun EU, 97–105. [https://www.juls.savba.sk/~vladob/papers/20171204_raslan_benko.pdf] pristupljeno 10.11.2024.
- Creese, S. 2017. Lexicographical Explorations of Neologisms in the Digital Age. Tracking New Words Online. Doktorska disertacija, Coventry University [<https://pu-reportal.coventry.ac.uk/en/studentTheses/lexicographical-explorations-of-neologisms-in-the-digital-age-tra>] Pristupljeno 10.11.2024.
- Fuertes-Olivera, P. 2012. Lexicography and the Internet as a (Re-)source. In: *Lexicographica - International Annual for Lexicography*, 49–70. DOI: 10.1515/le-xi.2012-0005
- Hundt, M., N. Nesselhauf, C. Biewer (eds.) 2007. Corpus Linguistics and the Web. *Language and Computers – Studies in Practical Linguistics*, 59.
- Kerremans, D., S. Stegmayr, H. Schmid 2012. The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In: *Current Methods in Historical Semantics*. Hrsg. von Kathryn Allan und Justyna Robinson. Berlin: Mouton de Gruyter, 59–96.
- Klosa-Kückelhaus, A., C. Tiberius 2024. 3 The Lexicographic Process. In: Klosa Kückelhaus, Annette (Ed.): *Internet lexicography. An introduction*. Berlin/Boston: de Gruyter, pp. 61–96. (Lexicographica : series maior 164) DOI: <https://doi.org/10.1515/9783111233758> [https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/12805/file/Klosa_Kueckelhaus_The_lexicographic_process_2024.pdf] Pristupljeno 10.11.2024.
- Krek, S. (ed.) 2024. *Book of Abstracts of the Workshop Large Language Models and Lexicography* 8 October 2024 Cavtat, Croatia [https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf#page=34] Приступљено 10.11.2024.
- Lazić Konjik, I., A. Milenković 2020. The Development of the Open Dictionary of Contemporary Serbian Language Using Crowdsourcing Techniques. In: Zoe Gavrilidou, Maria Mitsiaki, Asimakis Fliatouras (Eds.). *Proceedings of XIX Euralex Congress: Lexicography for inclusion*. Vol. I, Democritus University of Thrace, 479–484.
- Lenders, W. 2013. Computational lexicography and corpus linguistics until ca. 1970/1980. in: Gouws, R. H., Heid, U., Schweickard, W., Wiegand, H. E. (eds.) *Dictionaries - An International Encyclopedia of Lexicography*. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin: De Gruyter Mouton.
- Lüdeling, A., Evert, S. and Baroni, M. 2007. Using Web Data for Linguistic Purposes. *Language and Computers – Studies in Practical Linguistics*, 59, 14–31.
- Muhvić-Dimanovski, V. 2005. *Neologizmi, problemi teorije i primjene*. Zagreb: Filozofski fakultet, Zavod za lingvistiku.

- Prčić, T. 2019. *Engleski u srpskom*. Treće, elektronsko, izdanje. Novi Sad: Filozofski fakultet. [<https://digitalna.ff.uns.ac.rs/sadrzaj/2019/978-86-6065-512-9>] Pristupljeno 22.11.2024.
- Prčić, T. 2023. *Srpski sa engleskim. Osnove kontaktno-kontrastivne lingvistike*. Prvo, elektronsko, izdanje. Novi Sad: Filozofski fakultet. [<https://digitalna.ff.uns.ac.rs/sadrzaj/2023/978-86-6065-784-0>] Pristupljeno 22.11.2024.
- RNA: Vasić, V., T. Prčić, G. Nejgebauer. *Rečnik novijih anglicizama*. Novi Sad: Zmaj.
- SRNA: Prčić, T., J. Dračić, M. Milić i dr. *Srpski rečnik novijih anglicizama*. Prvo, elektronsko, izdanje. Novi Sad, Filozofski fakultet, 2021. [<https://digitalna.ff.uns.ac.rs/sites/default/files/db/books/978-86-6065-636-2.pdf>] Pristupljeno 10.11.2024.
- Šetka Čilić, I., J. Ilić Plauc 2021. Today's Usage of Neologisms in Social Media Communication. *Journal of Humanities and Social Sciences*, 6 (1 (14)): 115–140. DOI: 10.51558/2490-3647.2021.6.1.115
- Šipka, D. (1998). *Osnovi leksikologije i srodnih disciplina*. Novi Sad: Matica srpska.
- Würschinger, Q. 2021. Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter. *Frontiers in Artificial Intelligence*, 4. doi: 10.3389/frai.2021.648583 [<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.648583/full>] Pristupljeno 10.11.2024.
- Würschinger, Q., M. Fazleh Elahi, D. Zhekova, H. Schmid 2016. Using the Web and Social Media as Corpora for Monitoring the Spread of Neologisms. The case of rapefugee, rapeugee, and rapugee. Paul Cook, Stefan Evert, Roland Schäfer, Egon Stemle (eds.), *Proceedings of the 10th Web as Corpus Workshop*, 35–43. [<https://aclanthology.org/W16-2605/>] Pristupljeno 10.11.2024.

Ivana Lazić Konjik

Institute for the Serbian Language of the Serbian Academy of Sciences and Arts,
Belgrade

LEXICOGRAPHY IN THE DIGITAL ERA: INTERNET AND SOCIAL MEDIA AS SOURCES OF NEW WORDS

Summary

This paper examines the concept of the internet and social media as lexicographic corpora, considering the hypothesis that the internet serves as a rich repository of data that can be effectively utilized for lexicographic purposes. It describes the practices of a newly launched project for developing an online dictionary of Serbian neologisms (Neo21), which uses the internet as a corpus for collecting, selecting, and documenting new words, as well as for creating dictionary entries. The

goal of this dictionary is to provide users with quick and easy access to lexicographic data, offering the most up-to-date and reliable information along with precise explanations of meanings. The conclusion highlights that the internet and social media are crucial and contemporary sources for tracking and documenting linguistic changes, as well as a representative corpus for creating new types of dictionaries that monitor language development in real-time.

Keywords: internet, social media, lexicographic corpora, dictionaries of new words, Neo21, lexicography, Serbian language.