811.163.41'322.2
811.163.41'373
https://doi.org/10.18485/msc.2019.48.3.ch13

Бранислава Б. ШАНДРИХ
Филолошки факултет
Универзитета у Београду
Ранка М. СТАНКОВИЋ
Рударско-геолошки факултет
Универзитета у Београду
Мирјана С. ГОЧАНИН
Институт за српски језик САНУ

ЧИЈИ ЈЕ ПРИМЕР? АНАЛИЗА ЛЕКСИЧКИХ ОБЕЛЕЖЈА НА ПРИМЕРИМА РЕЧНИКА САНУ

У овом раду поставља се питање: да ли се може утврдити ко је аутор неког текста уколико се анализирају искључиво његова лексичка обележја? Како бисмо покушали да добијемо одговор на ово питање, посматрали смо примере у оквиру речничког чланка појединачне лексеме Речника САНУ, који су забележени у пет томова (и то: I, II, XVIII, XIX и XX). Сваки пример је преузет из неког извора на шта упућују скраћенице, наведене у заградама. Од преко 5.000 понуђених извора, определили смо се да ово истраживање спроведемо на дванаест извора односно на делима дванаест наших истакнутих књижевника, филолога и научника и то: на делима Јанка Веселиновића, Јована Јовановића Змаја, Борислава Пекића, Бранка Ћопића, Вука Стефановића Караџића, Иве Андрића, Милана Милићевића, Мирослава Крлеже, Петра Петровића Његоша и Стојана Новаковића. На датом материјалу издвојили смо различита лексичка обележја и визуелно их представили, а онда добијене резултате међусобно упоредили.

Кључне речи: идентификација ауторства, лексичка обележја, анализа обележја, примери Речника САНУ

1. Увод

Откривање аутора писаног текста није само модеран проблем. Неретко се аутори сакрију иза некаквог псеудонима и под другим именом објављују различите научне радове, књижевна дела, новинске чланке и слично. Помним проучавањем рукописа, начина изражавања, употребе интерпункције, врста речи, лексике и многих других елемената, често се анонимни или непотписани текстови могу довести у везу са својим правим творцем. Научна

дисциплина која се бави оваквим проучавањем текста назива се стилометрија и њен основни задатак јесте управо утврђивање ауторства.

Основним критеријумом за одређивање ауторства у писаним текстовима сматра се (нечији) рукопис. У куцаним и дигитално презентованим текстовима примена овог критеријума уопште није могућа. Рачунарска анализа, ипак, пружа могућност брзе обраде великог броја текстова, аутоматску упоредну анализу више дела истовремено и препознавање наизглед скривених правилности, које би људском оку могле промаћи.

Модерна математичко-рачунарска дисциплина *машинско учење* (енгл. Machine Learning) припада области вештачке интелигенције (енгл. Artifical Inteligence). Уопштено речено, машинско учење бави се подучавањем машина (рачунара, робота и сл.) да обављају разноврсне задатке својствене људима. Ово подучавање често се заснива управо на препознавању сакривених образаца и правилности у различитим типовима података те се још и назива „препознавање образаца" (енгл. Pattern Recognition). Подучавање рачунара да разуме и производи природан језик зове се обрада природних језика (енгл. Natural Language Processing). А обрада природних језика решава различите проблеме у разумевању текста који се тичу препознавања говора (енгл. Speech Recognition), сажимање текста (енгл. Text Summarization), лексичког рашчлањавања (енгл. Dependency Parsing), обележавања текста према врсти речи (енгл. Part-of-Speech Tagging), лематизације (енгл. Lemmatization), препознавања именованих ентитета (енгл. Named Entity Recognition), класификације текста (енгл. Text Classification) и многих других. Под класификовањем текста подразумева се обучавање математичког модела да аутоматски, на основу садржаја, тексту додели одговарајућу класу (категорију) или више њих, из скупа унапред договорених класа. Идентификација ауторства спада у домен класификације текста. У том случају, класе представљају могући аутори текста.

Шандрих (2018) анализира корпус кратких телефонских порука посматрајући сваку поруку као засебан текст, који са собом носи информације: о укупном броју карактера, о просечној дужини речи, о употреби интерпункције, о коришћењу малих и великих слова, броју, типу и учесталости емотограма и сл. У раду се поставља питање: да ли је могуће техникама машинског учења обучити математички модел (са опцијом за аутоматску идентификацију) да аутоматски идентификује аутора кратке поруке? Да би добио одговор на ово питање, аутор разматра две могућности. С једне стране, поруке бивају подељене у две групе: (1) поруке које је откуцао власник телефона из којег су и преузете и (2) све остале поруке. Обе групе заступљене су подједнако. С друге стране, поруке су класификоване на: (1) „аутоматски произведене", добијене од неког јавног сервиса (то су: обавештења из банке, уплате паркинг места, информације о такси возилу и сл.) и (2) личне поруке, које су састављали људи. Личне поруке су заступљеније од „аутоматски произведених".

У првом случају, имплицитно се испитује ауторство једне особе. То се чини пажљивим одабиром разноврсних лексичких обележја. У другом случају, разматрањем истог скупа обележја, испитују се правилности садржане у генерисаним порукама.

У овом раду истражујемо да ли је могуће пронаћи неке специфичности у примерима из дела појединих аутора, заступљених у Речнику САНУ. Анализирани су примери из дела дванаест истакнутих књижевника, филолога и научника и то: примери из дела Јанка Веселиновића, Јована Јовановића Змаја, Борислава Пекића, Бранка Ћопића, Вука Стефановића Караџића, Иве Андрића, Милана Милићевића, Мирослава Крлеже, Петра Петровића Његоша и Стојана Новаковића. Подстакнути анализом изложеном у раду Шандрих (2018), одлучили смо се да спроведемо сличан експеримент на датом материјалу. Обогатили смо избор обележја и упоређивали њихове вредности истовремено на примерима из дела одабраних аутора. Оваква анализа заправо представља корак ка изради математичког модела машинског учења за аутоматску идентификацију ауторства.

2. О Речнику САНУ

Речник српскохрватског књижевног и народног језика (скраћено Речник САНУ[1]) одређује се као описни, једнојезични лексикон тезаурусног типа. Наслов јасно сугерише да обрађена грађа, ексцерпирана из преко 5.000 извора[2] у временском периоду од готово два столећа, „покрива" домене књижевног и народног српског језичког ареала. Унапред успостављен теоријско-методолошки приступ у изради овог дела, креиран од стране наших највећих лексикографа[3] (иначе утемељивача и еминентних представника београдске лексикографске школе) и уопште истакнутих лингвиста, сврстава га у научну монографску публикацију од изузетног националног значаја[4]. А бројна лингвистичка истраживања[5] потврђују да се Речник САНУ може употребити као примарна односно изворна литература. Он презентује драгоцене податке из различитих области[6]. Тако се на основу етнографских студија, коришћених

[1] Описом и значајем Речника САНУ баве се радови: Вуловић–Ђинђић–Радоњић (2008), Ђинђић (2014), Ивановић (2013, 2017), Фекете (1993), а о важности Института за српски језик САНУ за српску лексикографију говори се у раду Ристић (2007).

[2] Стијовић (2017: 201) наводи да грађа за Речник САНУ има око 300 збирки речи из народних говора и да се бележи на око 6.000.000 листића.

[3] Теоријска и методолошка питања у изради Речника САНУ обрађивали су и разрешавали кроз радове Ирена Грицкат (в. Грицкат 1960а, 1960б, 1988а, 1988б), Даринка Гортан-Премк (в. Гортан-Премк 1980а, 1980б, 1982), Митар Пешикан (в. Пешикан 1963, 1966, 1967, 1973–1974, 1978, 1982), Милица Радовић-Тешић (в. Радовић-Тешић 1982) и Александар Белић у *Уводу* првог тома Речника САНУ (в. Белић 1959), а интерно Упутство, намењено превасходно основним обрађивачима у раду на Речнику САНУ, саставили су у највећој мери Ирена Грицкат, Митар Пешикан и Драгослав Илић. Упутство за израду Речника САНУ бива допуњено 2017. године.

[4] И *Закон о Речнику Српске академије наука и уметности,* објављен у Службеном гласнику РС бр. 110/05, у члану 1. сматра Речник САНУ подухватом „од изузетног значаја за националну културу и светску и домаћу науку" (в. http://www.mpn.gov.rs/wp-content/uploads/2015/08/zakon_o_recniku_sanu-cir.pdf).

[5] Навешћемо само нека језичка испитивања: Ђинђић (2009), Јакић (2015), Јовановић (2009), Лазић-Коњик (2009), Миланов (2017), Павковић (1984), Радоњић (2009).

[6] Набројаћемо неке од заступљених области и то: области спорта, политике, уметности, науке, културе, терминологије, технике и др.

у обради појединачних лексема, могу стећи знања о традиционалном начину обрађивања земљишта, о народним обичајима или веровањима, о разним (често данас ишчезлим) занимањима или о неким другим сферама живота наших предака. Објављивањем таквог садржаја Речник САНУ с једне стране показује и истиче вредност српске културне баштине, а с друге чува ово наслеђено благо од заборава.

Имајући у виду његову неспорну важност, припреми текста за електронско појављивање у јавност приступа се с нарочитом пажњом и одговорношћу. Кориговање текстуалног материјала за будуће дигитално издање[7] пет томова Речника САНУ (тј. првог, другог, осамнаестог, деветнаестог и двадесетог) тече у фазама. Изнећемо само нека запажања. Први постављени захтев односи се на то да оно што је забележено у штампаној верзији појединачног тома мора стајати и у електронској књизи. У току самог рада на добијеном тексту показује се да, зарад тачности, прецизности и јасноће садржаја који се представља, треба увести неке измене. И пошло се од једноставнијег ка сложенијем. Одлучено је тако да се, уместо техничке скраћенице И. која значи *исти извор,* користе потпуни подаци о извору односно да се уведе одговарајућа скраћеница за дати извор. У штампаним издањима речничких томова употреба овог И. везује се превасходно за уштеду простора. У штампаној верзији Речника САНУ (в. први том: **бȁч** и двадесети том: **плѐскāч**) бележи се: **бȁч** м (тур. baç) ист. *в. баждарина (3).* — Бач је једна стара дажбина која је узимана по кланцима (Гавр. М. 2, 420). То су били доходци београдске царинарнице, трошарине, кантара, бача на градским капијама (И., 377); **плѐскāч,** -а́ча м покр. експр. *онај који много и непромишљено говори, прича, брбљивац, блебетало;* исп. плескати и пљескати (I, 5). — Такога плескача тешко је наћ (Крња Јела, ЦГ, Вукс. М.). Велики је то плескач (И.). У дигиталном издању Речника САНУ требало би да стоји: **бȁч** м (тур. baç) ист. *в. баждарина (3).* — Бач је једна стара дажбина која је узимана по кланцима (Гавр. М. 2, 420). То су били доходци београдске царинарнице, трошарине, кантара, бача на градским капијама (Гавр. М. 2, 377); **плѐскāч,** -а́ча м покр. експр. *онај који много и непромишљено говори, прича, брбљивац, блебетало;* исп. плескати и пљескати (I, 5). — Такога плескача тешко је наћ (Крња Јела, ЦГ, Вукс. М.). Велики је то плескач (Крња Јела, ЦГ, Вукс. М.).

Детаљним прегледањем штампане верзије текста утврђено је да се експираторни акценат наводи на три начина: (1) у првом тому користи се ознака за краткоузлазни акценат односно од увођења фонтова Akademija, обележава се фонтом Akademija 01 (в. пример у оквиру речничког чланка лексеме **бȉчāк²**); (2) у деветнаестом тому употребљава се знак за дугоузлазни акценат односно од увођења фонтова Akademija, означава се фонтом Akademija 02 (в. примере у оквиру речничких чланака лексема **па́на², пȁнаȋр, пȁнаирȋште** и **панаȋрȋште**) и (3) бележи се новоустановљени, посебни знак за експираторни акценат успостављен коришћењем фонта Akademija 07 (в. приме-

[7] Питања дигитализације грађе за Речник САНУ и Речника САНУ проучавају радови: Ивановић–Јакић–Ристић (2016), Стијовић (2017), Стијовић–Станковић (2018).

ре у оквиру речничничких чланака лексема **пàздер** и **палàмāндра**). Треба напоменути да се знаковима за краткоузлазни или дугоузлазни акценат (у недостатку одговарајуће, посебне акценатске ознаке) заправо обележавало само место експираторног акцента, а не његов квалитет, нити квантитет. Будући да је, приликом преласка са фонтова COOC на фонтове Akademija, установљен посебан фонт за означавање поменутог акцента, требало би у електронској верзији Речника САНУ једнообразно бележити експираторни акценат коришћењем фонта Akademija 07. Складиштење у базу подразумева конвертовање акцената у одговарајуће комбиноване UNICODE карактере.

Пажљивим читањем првог и другог тома, а онда осамнаестог и деветнаестог тома штампаних издања Речника САНУ установило се да су неке техничке скраћенице употребљаване само у првим томовима, а касније су замењене другим. Тако се скраћеница св. која означава *свезу* појављује у првом (в. лексему **ако**) и другом тому (в. лексему **бȳдȳћ** и **бȳдȳћи** и **будȕћи**). У другом тому бележи се и скраћеница везн. којом се упућује на *везник* (в. лексему **већ**). Примери: (1) **ако** св. **1.** *погодбена, за означавање услова под којим се радња основног исказа има вршити* **а.** *под условом да, под погодбом да;* исп. аконо. … **2.** *временска, за означавање услова под којим се нешто увек дешава.* … **3.** *(ако, ако и) допусна, за означавање допуштања, претпостављање нечега као могућног и поред сметње: иако, мада, премда, макар да;* исп. акопрем, баш[2] (1д). … **4.** *намерна, за означавање онога што је циљ радње у основном исказу: да, еда, како, не би ли.* … **5.** *исказна, за означавање онога што је садржина основног исказа* **а.** *што, то што.* … **б.** *да ли;* (2) (а) **бȳдȳћ** и **бȳдȳћи** и **будȕћи** св. *узрочна, обично са свезом „да”: с обзиром на то што; пошто, јер.* (б) **већ**[1] везн.; вар. вет[1], веће **1.** *супротно* **а.** *после одричног исказа: него;* исп. а[3] (1ђ), венг, венго, већем, већен, већер, веш. … **2.** нар. *у поређењу: него;* исп. венгор, вендар. У осамнаестом (в. лексему **òткадгод** и **откàдгод**) и деветнаестом (в. лексему **па**[1]) тому искључиво се користи скраћеница везн.: (3) **òткадгод** и **откàдгод** везн. покр. *било откада, ма откада.*

У дигиталном издању Речника САНУ скраћеницу св. треба заменити скраћеницом везн. и пратећи текст у оквиру дефиниције прилагодити мушком роду (везн. погодбени … временски уместо св. погодбена … временска). Примери: (1) **ако** везн. **1.** *погодбени, за означавање услова под којим се радња основног исказа има вршити* **а.** *под условом да, под погодбом да;* исп. аконо. … **2.** *временски, за означавање услова под којим се нешто увек дешава.* … **3.** *(ако, ако и) допусни, за означавање допуштања, претпостављање нечега као могућног и поред сметње: иако, мада, премда, макар да;* исп. акопрем, баш[2] (1д). … **4.** *намерни, за означавање онога што је циљ радње у основном исказу: да, еда, како, не би ли.* … **5.** *исказни, за означавање онога што је садржина основног исказа* **а.** *што, то што.* … **б.** *да ли*; (2) (а) **бȳдȳћ** и **бȳдȳћи** и **будȕћи** везн. *узрочни, обично са везником „да”: с обзиром на то што; пошто, јер.*

Треба додати још и то да су се у првим томовима користиле скраћенице: (1) и. за обележавање *источног изговора;* касније је замењује скраћеница ек. којом се упућује на *екавски изговор*; (2) ј. за означавање *јужног изговора;* кас-

није је замењује скраћеница ијек. којом се упућује на *ијекавски изговор*; (3) з. за бележење *западног изговора*; касније је замењује скраћеница ик. којом се упућује на *икавски изговор*. Примери: (1) I: **бѐснети** ј. **бјѐсњети** (некњ. и. бѐснити ...); II: **бӯђавети** (буђáвети) ј. **бӯђавјети** (буђáвјети) (најчешће и. и ј. бӯђавити, буђáвити); **велѐлепан** ијек. **велѐљепан,** -пна, -пно (ређе ек. велѐљепан); XVIII: **отмењàкуша** (ек.); XIX: **òцелан** ијек. **òцјелан,** -лна, -лно (обично одр. òцелнӣ, -лнā, -лнō; ретко ек.); XX: **пешàдāр,** -áра м (ек.); (2) I: **àвијест**[1] ж (само ј.); II: **бурѐвеснӣк** ј. **бурѐвјеснӣк**; **вàвēст** ијек. **вàвијест**; XVIII: **опопре́чити се,** – òпрēчӣм се ијек. **опопријѐчити се,** -òпријечӣм се; XIX: **оцѐнилац** ијек. **оцјѐнилац**; XX: **пешàдија** ијек. **пјешàдија**; (3) I: **биљ-** (ј. и з.); II: **била буника** (само з.); **вӣдиочев** (ијек. и ик. и некњ. ек.); XVIII: **òпрēд(а)** ијек. **òприjед(а)** (ик. оприд); XIX: **òцерити се,** -ӣм се (ек. и ијек.) (дијал. ијек. оцјерити се; ик. оцирити се); XX: **пӣвалӣште** (ик.). Примећује се да се у првом тому и првом делу другог тома доследно употребљавају првобитно успостављене скраћенице: и., ј. и з. У другом делу другог тома, осамнаестом, деветнаестом и двадесетом тому користе се скраћенице: ек., ијек. и ик. Потребно је првобитне заменити новим скраћеницама на свим местима где се појављују.

Предложак текста за електронску верзију Речника САНУ у поменутим томовима треба кориговати и тако да се: (1) не употребљава латинична графема у ћирилицом написаној речи и обрнуто; (2) не користи графема из симбола са наводном ознаком дужине на месту где треба употребити фонт Akademija 05; (3) не појави нетачни редни број уз одређену семантичку реализацију.

3. Коришћени текстуални ресурси

Примена савремених технологија и дигитализација свих томова Речника САНУ би омогућила да се он константно допуњава и осавремењује. Дигитализовани томови овог лексикона могли би послужити као корпус у раду на будућим томовима, а тако би се и процес израде Речника САНУ знатно убрзао.

Први резултати рада на дигитализацији, која је отпочела 2016. године, публиковани су у раду Стијовић/Станковић (2018) где су дате: процедуре за анализу дигитализованог текста 1. и 19. тома, сегментирани речнички чланци и информационе целине у оквиру речничког чланка према моделу. У наставку рада на дигитализацији обухваћени су и други томови, што је омогућило и анализу коју приказујемо у овом раду.

До сада је кориговано и импортовано у базу 5 томова и то: I, II, XVIII, XIX и XX, што представља скуп података за истраживање, презентовано у овом раду. Како се и даље интензивно ради на кориговању текста осталих томова, верујемо да ће нека будућа истраживања моћи детаљније да сагледају концепте анализе дате у овом раду.

Скуп података из пет томова Речника САНУ садржи: ~ 60.000 речничких чланака са ~ 105.000 лексичких јединица (значења или семантичких реализација). Око 11.500 речничких чланака има више од једне лексичке јединице односно има нумерисане лексичке јединице (разграната значења).

Приликом ексцерпције примера, редактори могу да интервенишу на два начина: (1) могу да изоставе део оригиналне реченице, што се обележава симболом „ …” и (2) могу да унесу део реченице у угласте заграде уколико је потребно нешто додатно појаснити.

На основу анализираног скупа података, утврђујемо да се велики број примера не скраћује (чак 71% од укупног броја обрађених примера). Код извесног броја примера (22% од укупног броја обрађених примера) бележи се једном симбол „ …”, незнатан број примера има два пута забележен овај симбол (6% од укупног броја обрађених примера), а само мали број примера има наведен овај симбол више од два пута (1% од укупног броја обрађених примера). Мали број примера има унету реч или део реченице у угласте заграде (само 7% од укупног броја обрађених примера). То значи да је велики број примера наведен без уметнутог текста у угласте заграде (чак 93% од укупног броја обрађених примера). Ако сагледавамо скуп обрађених примера с обзиром на то да ли је пример забележен без редакторских интервенција или са неком од две и(ли) обе евидентиране и горе наведене, онда се показује следеће: (1) знатан број цитираних примера није имао редакторске интервенције (66% од укупног броја обрађених примера), известан број примера има једном забележен симбол „ …” и нема уметнутог текстуалног садржаја у угластим заградама (20% од укупног броја обрађених примера), невелики број примера има употребљен симбол „ …” два пута или више пута, а нема забележног уноса текста у угласте заграде (6% од укупног броја обрађених примера), незнатан број примера има евидентиран само унос речи или дела реченице у угласте заграде (5% од укупног броја обрађених примера) и само код малог броја примера употребљен је и симбол „ …” и унесен је текстуални материјал у угласте заграде (2% од укупног броја обрађених примера). Треба напоменути да је велики број речничких чланака појединачних лексема имао цитиране примере (чак 70% од укупног броја речничких чланака).

4. Веб сервис и израчунавање обележја

Процедура екстракције обележја тече на следећи начин: рачунарски програм израчунава нумеричке вредности различитих обележја за сваки пример (то важи и за произвољан текст). Потом текст добија такозвану векторску репрезентацију. Она се може употребљавати приликом обучавања математичких модела.

Размотримо конкретан пример: „Ко рано рани, две среће граби!” Узевши у обзир следећа обележја: (1) укупан број карактера, (2) укупан број малих слова, (3) укупан број великих слова, (4) укупан број узвичника и (5) укупан број цифара, векторска репрезентација овог текста редом изгледала би: (30,

22, 1, 1, 0). То значи да дати текст има укупно тридесет карактера, двадесет два мала слова, једно велико слово, један узвичник и нема забележених цифара. Дата векторска репрезентација може се даље користити у неким израчунавањима.

Ради аутоматског произвођења примера у оквиру речничког чланка, Едер и др. (2016), Килгариф и др. (2008) и Косем (2017) у својим радовима разматрају различита лексичка обележја. Дата истраживања утицала су на одабир обележја која су сагледавана у нашем раду.

За потребе овог, али и других истраживања која се заснивају на вредностима оваквих обележја, развијен је алат за аутоматско добијање векторске репрезентације текстова. Реч је о веб сервису, коме се, као улаз, зада текст. Као излаз добијају се парови: обележје − вредност. Алат израчунава укупно 41 лексичко обележје. Издвојићемо и груписаћемо нека:

- на нивоу карактера:
  - sentence_length: укупан број карактера;
  - no_digits: укупан број цифара;
  - no_weird_chars: укупан број специјалних карактера из групе („#$%&\’()*+-/:;<=>?@[\\]^_`{|}~’„”…);
  - no_commas: укупан број размака;
  - no_punctuation: укупан број знакова интерпункције;
- на нивоу речи:
  - no_all_tokens: укупан број токена;
  - avg_token_len: просечна дужина токена;
  - max_token_len: дужина најдужег токена;
  - no_all_words: укупан број речи;
  - avg_word_len: просечна дужина речи;
  - no_capitalised_words: укупан број речи са почетним великим словом, а не налазе се на почетку текста;
  - no_rare_tokens: број токена са фреквенцијом мањом од неке задате вредности у референтном корпусу;
  - Avg_freq_in_corpus: просечна фреквенција речи присутних у тексту у односу на референтни корпус;
- остало:
  - no_pronouns: укупан број личних заменица.

Навешћемо као пример употребу веб сервиса за екстракцију ових обележја помоћу програмског алата *curl*:

```
curl -d '{
"data": "Ко рано рани, две среће граби!",
"feature_names": ["sentence_length","no_punctuation", "no_all_tokens"]
}'
-H "Content-Type: application/json" -X POST http://147.91.183.8:12347/features
```

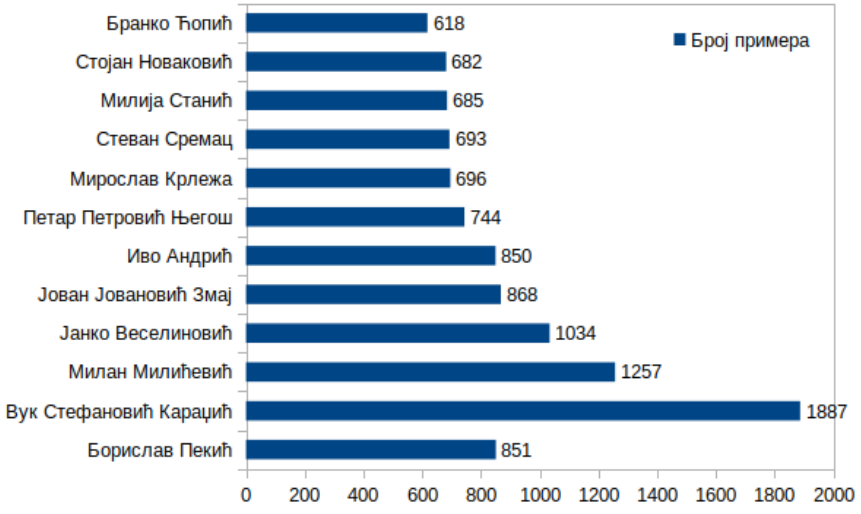
Као вредност поља *data* бележи се текст, као вредност поља *feature_names* наводи се листа назива обележја која се рачунају за дати текст. Уколико се то поље изостави, сервис враћа вредности свих обележја које систем у датом тренутку подржава. У конкретном примеру излаз је:

```
{"sentence_length": 30, "no_punctuation": 2, "no_all_tokens": 8}
```

У току је и израда интерфејса за веб апликацију[8] која ће омогућити унос текста у поље предвиђено за то, а као одговор дати вредности одабраних обележја.

5. Упоредна анализа лексичких обележја

На слици 1. дат је тачан број примера из дела наших дванаест изузетних књижевника, филолога и научника, цитираних у обрађеним томовима Речника САНУ. Уочава се да је убедљиво највећи број примера у оквиру речничког чланка појединачне лексеме наведен из дела Вука Стефановића Караџића.



Слика 1. Број примера из дела наведених аутора

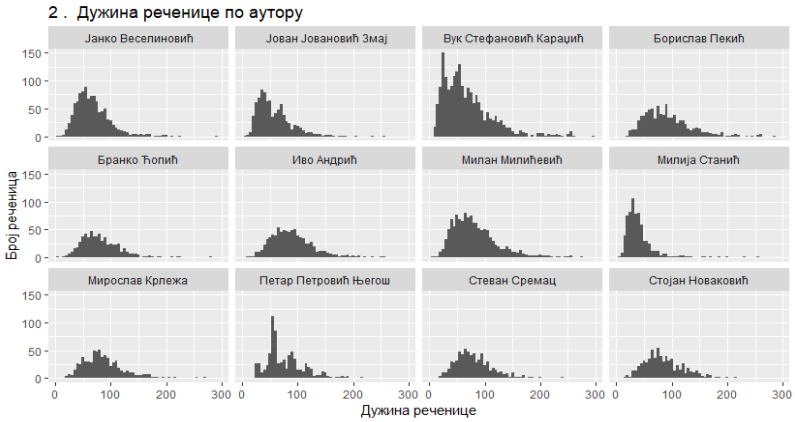
На слици 2. приказани су хистограми расподеле дужине примера, исказани у броју карактера. На апсциси је наведен број карактера, а на ординати број примера који имају одговарајућу дужину. Може се утврдити да је расподела на основу примера из дела Иве Андрића и Бранка Ћопића прилично

[8] Веб интерфејс који је у изради инспирисан је интерфејсом датим на адреси http://features.jerteh.rs/

нормална и симетрична, а ако се посматрају примери из дела Вука Караџића и примери Милије Станића, хистограм се нагомилава улево. То значи да су примери махом мање дужине.

За анализу скупова података потребно је обично више информација него што је потребно за приказивање мера централне тенденције односно за приказивање средње вредности, моде (као најучесталије вредности) и медијане. Неопходно је имати информације о варијабилности или дисперзији података. За визуелизацију података користи се boxplot график.

Кутијасти дијаграм (који се још назива и правоугаони или дијаграм са брковима; на енглеском: boxplot, box and whisker, B&W) је стандардизовани начин приказивања дистрибуције података на основу пет вредности („минимум", први квартил (К1), медијана, трећи квартил (К3) и „максимум"). Квартили деле скуп података на четири једнака дела. Сваки део садржи четвртину од укупног броја података односно 25%. Дијаграм јасно детектује екстреме (outliers) и њихове вредности. Он приказује да ли су испитивани подаци симетрични, колико су чврсто груписани и да ли су тачни или не. Изгледа врло једноставно у поређењу са хистограмом или дијаграмом густине. Заузима мање простора, што представља предност у случајевима када се међусобно пореде дистрибуције више група или скупова података.



Слика 2. Хистограми расподеле дужине примера исказани у броју карактера

На графицима се обично обележава 5 кључних вредности:

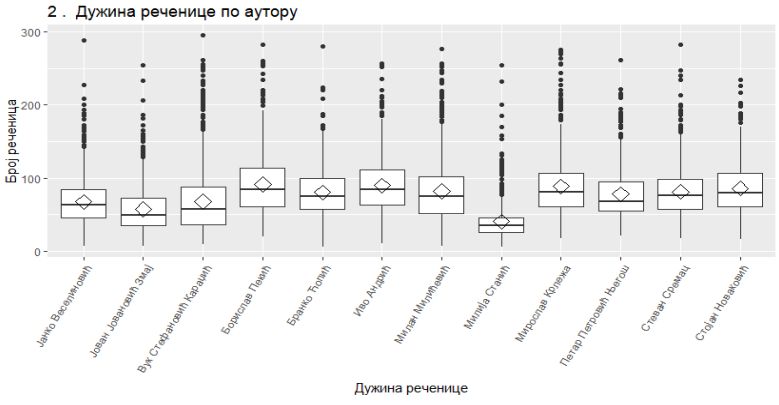
- медијана (Q2 / 50-ти перцентил): средишња вредност у скупу података који су уређени од најмањег ка највећем; број који раздваја горњу половину узорка (средина кутије);
- први квартил (Q1 / 25-ти перцентил): средњи број између најмањег броја (не „минимум") и медијане скупа података (доња ивица кутије кутије);

- трећи квартил (Q3 / 75-ти перцентил): средња вредност између медијане и највеће вредности (не „максимума") скупа података (горња ивица кутије);
- интерквартилни опсег (IQR): од 25-ог до 75-ог перцентила;
- бркови екстремне вредности (приказане као тачке);
- „Максимум": Q3 + 1.5 * IQR (ово није стварни максимум и зато је обележен наводницима);
- „Минимум": Q1 -1.5 * IQR (ово није стварни минимум и зато је обележен наводницима).

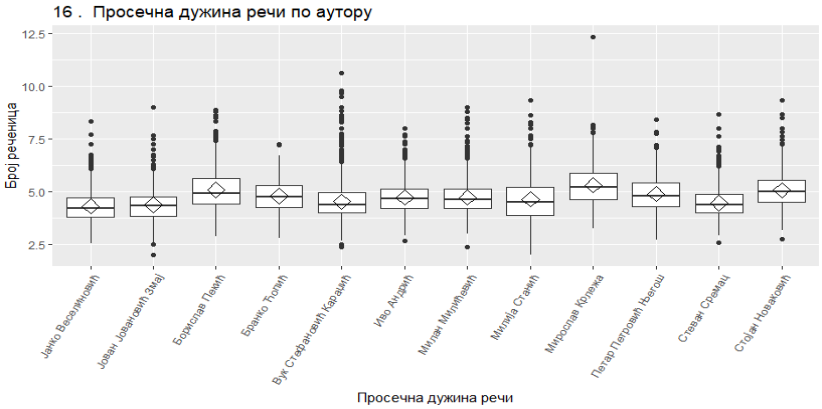
Кутијасти дијаграми представљају један од начина приказа статистичких параметара у овом раду. То се показује као логичан избор будући да је наш задатак био да упоредимо примере из дела истакнутих књижевника, филолога и научника према различитим обележјима.

На слици 3. уочава се да су примери из дела Бориславa Пекића и Иве Андрића дужи и укупно имају знатно већи број карактера у односу на примере из дела других аутора. Примери Милије Станића убедљиво су најкраћи.

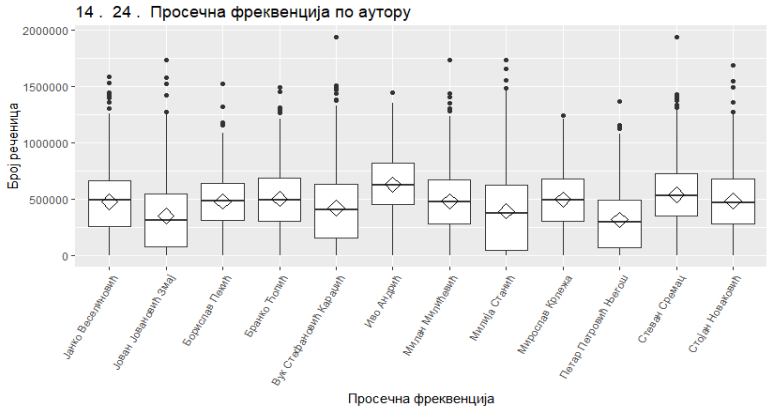
Један од параметара који смо израчунавали јесте просечна дужина речи. Анализирани резултати приказани су на слици 4. Видљиво је да Мирослав Крлежа и Борислав Пекић користе дуже речи. Да бисмо проценили колико су речи које се користе у примерима познате читаоцу, одабране су фреквенције из корпуса СрпКор 2013 (в. Утвић 2014). Свакој речи придружена је фреквенција из корпуса, а онда је израчуната средња вредност. На слици 5. може се уочити да Иво Андрић у својим делима користи веома фреквентне речи. Примери Милије Станића имају највећу дисперзију фреквентности. За примере из дела Петра Петровића Његоша и Јована Јовановића Змаја карактеристично је да имају најниже медијане. У даљим анализама ће се опционо елиминисати функционалне речи из рачунања средње вредности, што ће можда показати изражене разлике.



Слика 3. Кутијасти дијаграми дужина реченица



Слика 4. Кутијасти дијаграми просечне дужине речи



Слика 5. Просечна фреквенција речи у односу на референтни корпус СрпКор2013

6. Визуелно представљање текстова у простору

Пре конструисања било каквог математичког модела, потребно је познавати податке које треба обрадити. Зато смо желели да нашу колекцију докумената представимо визуелно. Текстови су претходно обрађени на следећи начин: груписани примери лематизовани су коришћењем тагера (в. Утвић 2014), а затим је примењена листа стоп речи. Коришћењем електронских речника и алата Unitex (в. Витас и Крстев 2012) елиминисане су све врсте речи изузев именица.

6.1. Моделирање преовлађујућих тема

Моделирање преовлађујућих тема у тексту постала је популарна процедура за груписање докумената у семантичке групе. Ова процедура омогућава да се од сирових текстуалних података добије интерактивна визуализација тематског модела, која за циљ има откривање тематских мотива, преовлађујућих у тексту који се разматра. Употребом алата TopicsExplorer произвели смо визуелизацију на основу најзаступљенијих тема односно концепата, које се помињу у примерима из дела одабраних аутора.

Topics Explorer је алат у ком је имплементиран алгоритам за латентну Дирихлеову алокацију (енгл. Latent Dirichlet Allocation), засновану на претпоставци да документи неке колекције садрже заједнички скуп латентних (скривених) тема у различитим односима чинећи Дирихлеову расподелу (в. Блеј и др. 2003). Број потенцијално скривених заједничких тема задаје се унапред. Тема се дефинише као мултиноминална расподела токена (речи, н-грама, концепата итд.) из утврђеног речника. То значи да ће се сваки токен јављати у свакој теми са одређеном вероватноћом појављивања. Све заступљене речи додељују се итеративним поступком свим темама. За сваку реч израчунава се вероватноћа да се та реч нађе у датој теми.

Добијена визуелна репрезентација приказана је на слици 6. Теме су представљене у колонама, а одабрани аутори у редовима. Тамнијом бојом указује се на већу вероватноћу да се нека тема појави у примерима из дела неког аутора. Теме су у излазним табелама и графичком приказу прдстављене трима речима. Уочава се да су човек, ствар и година концепти који преовлађују у примерима из дела Борислава Пекића, турчин, свијет и сила у примерима из Његошевих дела, земља, коњ, људи, пас, цар и брат у примерима из Вукових дела, господин, град и слика у примерима из Крлежиних дела, човек, жена и реч у делима Милана Милићевића, Змаја и Јанка Веселиновића, жена, човјек и вода у делима Иве Андрића и Бранка Ћопића, во, овца и новац у примерима Милије Станића, кућа, очи и рука у примерима из дела Јанка Веселиновића, живот, краљ и Србија у делима Стојана Новаковића итд.

6.2. Анализа главних компоненти

Анализа главних компоненти (енгл. Principal Component Analysis) је статистичка процедура која, користећи линеарне трансформације, текст, представљен као вектор, трансформише односно мапира у простор нижих димензија (в. Џолиф 2011). Такав поступак назива се пројекција и обавља се статистичким методама за идентификацију обележја којима се најбоље детерминише текст и(ли) описују највише варијабилности. У овом случај, не говоримо о обележјима која смо ми одабрали, већ о неким генеричким обележјима којима се текст може представити (скуп речи, скуп н-грама итд).

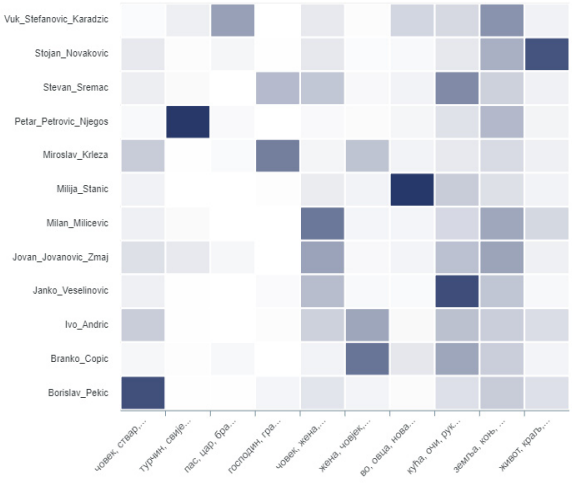
Слика 6. презентује примену овог модела на наш материјал. Приказује се пројекција примера из дела одабраних аутора на дводимензиони векторски простор. Распознају се четири највеће групе које представљају: (1) примере из дела Стојана Новаковића, Милана Милићевића, Борислава Пекића, Сте-

вана Сремца и Јанка Веселиновића; (2) примере из дела Мирослава Крлеже, Вука Стефановића Караџића, Јована Јовановића Змаја и Бранка Ћопић; (3) примере Милије Станића и (4) примере из дела Петра Петровића Његоша.

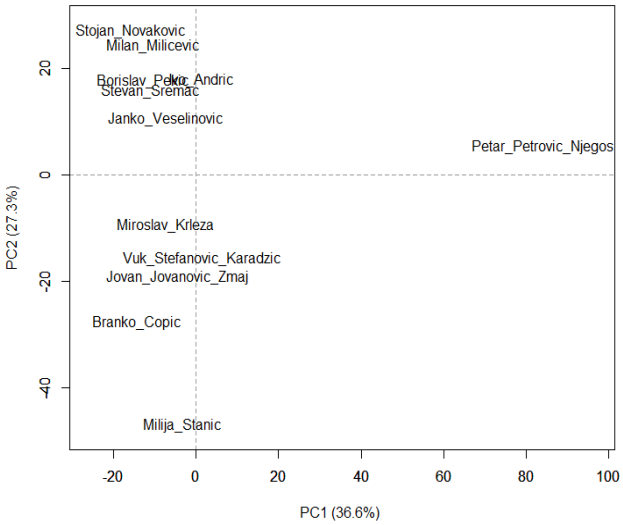
7. Закључци

Речник САНУ, користећи се фондом од преко 5.000 извора, представља праву ризницу пажљиво одабраних примера у оквиру речничког чланка појединачне лексеме. Овим радом желели смо да истражимо да ли је, применом математичко-рачунарске анализе на одабрани скуп, састављен од примера из дела дванаест истакнутих књижевника, филолога и научника, могуће установити неке ауторске специфичности. Најпре су одабрана обележја визуелно представљена на разне начине, а потом је и сваки резултат укратко коментарисан. Служећи се двема методама, посматрана колекција пројектована је тако да се документи (један документ је скуп свих примера из дела једног аутора) могу заједно приказати. Таквом анализом стекли смо више знања о садржају наше колекције и о односима међу документима.

Следећи задатак је обучавање модела машинског учења. То значи да ће се за неки пример, са одређеном вероватноћом, моћи одредити порекло односно утврдити ауторство. Зато анализом треба обухватити већи број извора.



Слика 6. Моделирање преовлађујућих тема



Слика 7. Анализа главних компоненти

ЛИТЕРАТУРА

Блеј и др. 2003: D. M. Blei, A. Y. Ng & M. I. Jordan, Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Белић 1959: Александар Белић, „Увод”, *Речник српскохрватског књижевног и народног језика САНУ,* Београд: САНУ, VII−XXVI.

Витас/Крстев 2012: Duško Vitas & Cvetana Krstev, Processing of Corpora of Serbian Using Electronic Dictionaries. Prace Filologiczne, vol. LXIII, Warszawa, 279–292.

Вуловић/Ћинћић/Радоњић 2008: Наташа Вуловић, Марија Ћинћић и Данијела Радоњић, „Реч више о Речнику САНУ”, *Књижевност и језик* LV/1–2, 171–179.

Гортан-Премк 1980а: Даринка Гортан-Премк, О граматичкој информацији и семантичкој идентификацији у великим описним речницима, *Наш језик*, XXIV/3, 107–114.

Гортан-Премк 1980б: Даринка Гортан-Премк, „О неким проблемима синтаксичке норме у Речнику САНУ”, *Научни састанак слависта у Вукове дане*, 10/1, 91–95.

Гортан-Премк 1982: Даринка Гортан-Премк, „Синонимски низ у лексикографској дефиницији”, *Научни састанак слависта у Вукове дане*, 12/1, 45–50.

Грицкат 1960а: Ирена Грицкат-Радуловић, „Академијски речници и њихови задаци – поводом прве књиге Речника Српске академије наука”, *Наш језик*, X/3–6, 88–100.

Грицкат 1960б: Ирена Грицкат-Радуловић, „Академијски речници и њихови задаци – поводом прве књиге Речника Српске академије наука", *Наш језик*, X/7–10, 212–217.

Грицкат 1988а: Ирена Грицкат-Радуловић, „Проблеми описне лексикографије", *Глас САНУ*, CCCLII, 13, 7–13.

Грицкат 1988б: Ирена Грицкат-Радуловић, „Речник Српске академије наука и уметности", *Глас САНУ*, CCCLII, 13, 25–40.

Ђинђић 2009: Марија Ђинђић, „О покрајинској лексици турског порекла у Речнику САНУ", *Дијалекат – дијалекатска књижевност: зборник радова*, Лесковац: Лесковачки културни центар, 110–114.

Ђинђић 2014: Марија Ђинђић, „Деветнаести том Речника САНУ", *Наш језик*, XLV/3–4, Београд: Институт за српски језик САНУ, 115–119.

Едер и др. 2016: M. Eder, J. Rybicki & M. Kestemont, Stylometry with R: a package for computational text analysis, R Journal 8(1): 107–121.

Закон о Речнику Српске академије наука и уметности: http://www.mpn.gov.rs/wp-content/uploads/2015/08/zakon_o_recniku_sanu-cir.pdf

Ивановић 2013: Ненад Ивановић, *Речник САНУ и његова улога у лексичкој стандардизацији српског језика (са историјског и лексикографског аспекта)* (необјављена докторска дисертација, одбрањена на Филолошком факултету Универзитета у Београду).

Ивановић–Јакић–Ристић 2016: Ненад Ивановић, Милена Јакић и Стана Ристић, „Грађа Речника САНУ – потребе и могућности дигитализације у светлу савремених приступа", у: *Лексикологија и лексикографија у светлу савремених приступа* (ур.Стана Ристић), Београд: Институт за српски језик САНУ, 133–154.

Ивановић 2017: Ненад Ивановић, „Један прилог историји српске лексикографије ('Рукописна збирка народних речи' В. С. Караџића у грађи за Речник САНУ)", *Наш језик*, XLVIII/3–4, 57–65.

Јакић 2015: Милена Јакић, „Формално означавање антонимије у Речнику САНУ", *Зборник Матице српске за филологију и лингвистику*, LVIII/1, 155–178.

Јовановић 2009: Владан Јовановић, „О језичком корпусу Речника САНУ и дијалекатској лексици", у: *Дијалекат – дијалекатска књижевност: зборник радова*, Лесковац: Лесковачки културни центар, 115–120.

Килгариф и др. 2008: A. Kilgarriff, M. Husák, K. McAdam, M. Rundell & P. Rychlý, GDEX: Automatically Finding Good Dictionary Examples in a Corpus, In E. Bernal & J. DeCesaris (eds.). Proceedings of the XIII EURALEX International Congress, Barcelona: Universitat Pompeu Fabra, 425–432.

Косем 2017: Iztok Kosem, Dictionary examples, In Dictionary of Modern Slovene: Problems and Solutions, V. Gorjanc, P. Gantar, I. Kosem, S. Krek (eds.). Ljubljana: University of Ljubljana, Faculty of Arts.

Косем и др. 2018: Iztok Kosem, Kristina Koppel, Tanara Zingano Kuhn, Jan Michelfeit & Carole Tiberius, Identification and Automatic Extraction of Good Dictionary Examples: the Case(s) of GDEX, International Journal of Lexicography.

Лазић-Коњик 2009: Ивана Лазић-Коњик, „О лексици народних говора у Речнику САНУ", у: *Дијалекат − дијалекатска књижевност: зборник радова*, Лесковац: Лесковачки културни центар, 157−161.

Миланов 2017: Наташа Миланов, *Полисемија српске лексике на корпусу Речника српскохрватског књижевног и народног језика САНУ* (необјављена докторска дисертација, одбрањена на Филолошком факултету Универзитета у Београду).

Павковић 1984: Васа Павковић, „Природа новоексцерпиране грађе за Речник САНУ", у: *Лексикологија и лексикографија: зборник радова*, Нови Сад − Београд: Матица српска − Институт за српскохрватски језик, 109–115.

Пешикан 1963: Митар Пешикан, „О Речнику Српске академије наука и уметности", *Наш језик*, XIII/3–5, 169–196.

Пешикан 1966: Митар Пешикан, „О начелима обраде и развијања стручне терминологије", *Наш језик*, XV, 180–194.

Пешикан 1967: Митар Пешикан, „Невоље рада на нашим описним речницима", *Наш језик*, XVI, 193–204.

Пешикан 1973–1974: Митар Пешикан, „Трећина посла на изради Речника САНУ", *Наш језик*, XX/1–5, 11–22.

Пешикан 1978: Митар Пешикан, „Десет томова Речника САНУ", *Наш језик*, XXIII/3–4, 87–92.

Пешикан 1982: Митар Пешикан, „О селекцији речи у описним речницима", у: *Лексикологија и лексикографија: зборник реферата*, Београд – Нови Сад: САНУ и др., 209–215.

Радовић-Тешић 1982: Милица Радовић-Тешић, „Проблеми обраде фигуративних значења у описним речницима", у: *Лексикологија и лексикографија: зборник реферата*, Београд – Нови Сад: САНУ и др., 257–262.

Радоњић 2009: Данијела Радоњић, „Лексика лесковачког краја у Речнику Српске академије наука и уметности", у: *Дијалекат – дијалекатска књижевност: зборник радова*, Лесковац: Лесковачки културни центар, 243–246.

Речник САНУ: *Речник српскохрватског књижевног и народног језика*, I–XX, Београд: САНУ – Институт за српски језик САНУ, 1959–2017.

Ристић 2007: Стана Ристић, „Прва лексикографска школа у Институту за српски језик САНУ", у: *Шездесет година Института за српски језик САНУ: зборник радова I*, Београд: Институт за српски језик САНУ, 131–149.

Стијовић 2017: Рада Стијовић, „Грађа за Речник САНУ – благо које треба сачувати (о дигитализацији листића)", *Наш језик*, XLVIII/3–4, 201–207.

Стијовић/Станковић 2018: Рада Стијовић и Ранка Станковић, „Дигитално издање Речника САНУ: формални опис микроструктуре Речника САНУ", *Научни састанак слависта у Вукове дане*, 47/1, 427–440.

Упутство: *Упутство за израду Речника САНУ*, Институт за српск(охрватск)и језик САНУ (рукопис) 1959. и (допуњено) 2017.

Утвић 2014: Miloš Utvić, The construction of reference corpus of contemporary Serbian [Izgradnja referentnog korpusa savremenog srpskog jezika] (Doctoral dissertation, University of Belgrade).

Фекете 1993: Егон Фекете, „О Речнику српскохрватског књижевног и народног језика САНУ", у: *Сто година лексикографског рада у САНУ,* Београд: САНУ − Институт за српски језик САНУ, 21–49.

Џолиф 2011: I Jolliffe, Principal component analysis, Springer Berlin Heidelberg, 1094–1096.

Шандрих 2018: Branislava Šandrih, Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting, In 3rd International Conference Computational Linguistics in Bulgaria (CLIB 2018), Department of Computational Linguistics at the Institute for Bulgarian Language with the Bulgarian Academy of Sciences, 203–210.

Branislava B. Šandrih, Ranka M. Stanković, Mirjana S. Gočanin

WHOSE EXAMPLE IS IT? FEATURE ANALYSIS OF EXAMPLES FROM DICTIONARY OF SERBIAN ACADEMY OF SCIENCE AND ARTS

Summary

The question we ask ourselves in this paper is the following: Is it possible to determine who is the author of a text by analyzing various lexical features? In order to try to get an answer, we observed examples that support lexical entries listed in five of the total of twenty volumes of the Dictionary of Serbian Academy of Science and Arts. Each dictionary example is documented with its author, so we decided to examine only examples that origin from twelve great names in the domestic literature. For each author's example, we extracted different lexical features, and then we visualized and compared these results using different statistical methods.