# Импресум

ЗА ИЗДАВАЧА:
**проф. др Александар Јерков** office@unilib.rs
*Универзитетска библиотека „Светозар Марковић“*
*Филолошки факултет, Универзитет у Београду*

ИЗДАВАЧ:
*Филолошки факултет, Универзитет у Београду*
*Универзитетска библиотека „Светозар Марковић“*
*Заједница библиотека универзитета у Србији*

ГЛАВНИ И ОДГОВОРНИ УРЕДНИК:
**проф. др Цветана Крстев** cvetana@matf.bg.ac.rs

ТЕХНИЧКИ УРЕДНИК:
**др Александра Тртовац** aleksandra@unilib.rs
*Универзитетска библиотека „Светозар Марковић“*

УРЕДНИК ЕЛЕКТРОНСКОГ ИЗДАЊА:
**др Јелена Гавриловић** andonovski@unilib.rs
*Универзитетска библиотека „Светозар Марковић“*

ISSN 1450-9687 (штампано издање)
ISSN 2217-9461 (е-издање)

Београд, год. 23, бр. 1, јул 2023.

ИЗРАДА ВЕБ ПОРТАЛА ЧАСОПИСА:
**др Јелена Гавриловић**
*Универзитетска библиотека „Светозар Марковић“*

ЛЕКТОР ЗА СРПСКИ ЈЕЗИК:
**Свјетлана Ђелић**
*Универзитетска библиотека „Светозар Марковић“*

ДИЗАЈН И ПРИПРЕМА ЗА ШТАМПУ:
**Инфо*тека* тим**

РЕДАКТОР БИБЛИОГРАФИЈЕ И УДК КЛАСИФИКАЦИЈА:
**др Наташа Дакић**
*Универзитетска библиотека „Светозар Марковић“*

РЕДАКТОР ДОИ БРОЈЕВА:
**др Милош Утвић**
*Филолошки факултет, Универзитет у Београду*

РЕДАКЦИЈА ЧАСОПИСА:
**Часопис Инфо*тека***
*Булевар краља Александра 71, 11000 Београд*

+381 11 3370–211
infothecajdh@gmail.com

ШТАМПА:
**Мамиго плус**
*Београд*

*Штампање овог броја подржали су Друштво за језичке ресурсе и технологије JeРТех и Викимедија Србије у оквиру пројекта “Википодаци о српским романима (унос, визуелизација и анализа)”*

Часопис излази два пута годишње

# Садржај

# Аутоматска процена кратких одговора коришћењем латентне семантичке анализе

УДК 81'322.4

Теодора Михајлов
teodoramihajlov@gmail.com
*Универзитет у Београду*
*Београд, Србија*

**САЖЕТАК:** Употреба технологије у настави представља изазов како за предаваче, тако и за ученике. У овом раду, представили смо систем за аутоматску процену тачности одговора коришћењем Латентне семантичке анализе (LSA) – методе који почива на претпоставци да ће се речи сличног значења јавити у сличним контекстима. Систем ће бити коришћен у оквиру дигиталних лексичких картица за усвајање вокабулара другог језика, које ће пратити CLIL метод наставе. Резултати приказани у овом раду указују на то да LSA даје боље резултате на дужем тексту него на краћем који има форму дефиниције. Одговори су класификовани у две и у више категорија коришћењем KNN алгоритма. Процењена прецизност модела за две категорије је $P = 0,73$, одзив $R = 1,00$, а скор $F_1 = 0,85$, док код више категорија је добијено $P = 0,50$, $R = 0,47$, а мера $F_1 = 0,46$. Резултате је потребно узети са одређеном резервом због малог скупа над којим је вршено обучавање и евалуација.

**КЉУЧНЕ РЕЧИ:** LSA, CLIL, усвајање вокабулара другог језика, косинусна сличност, KNN

**РАД ПРИМЉЕН:** 1. април 2023.
**РАД ПРИХВАЋЕН:** 7. мај 2023.

## 1. Увод

Коришћење технологије у сврхе побољшавања наставе језика и исхода учења актуелан је проблем још од 60-их година 20. века. У овом раду бавићемо се израдом модела за аутоматску процену тачности

одговора коришћењем латентне семантичке анализе (енгл. *Latent Semanitc Analysis*, LSA). Модел ће касније бити имплементиран у оквиру дигиталних лексичких картица за учење вокабулара енглеског као страног језика.

Ранијим истраживањима (Landauer, Foltz, and Darrell 1998; Lemaire and Dessus 2003; Lifchitz, Jhean-Larose, and Denhière 2009), закључено је да LSA на добар начин представља велики број когнитивних способности човека, међу којима је и усвајање вокабулара, и да има висок степен слагања са проценама тачности евалуатора (Landauer et al. 1997; Graesser et al. 2000; Lemaire and Dessus 2003; Landauer, Laham, and Foltz 2003; Picca, Jaccard, and Eberlé 2015). Дигиталне лексичке картице показале су се као добро средство за учење вокабулара другог језика (Ashcroft, Cvitkovic, and Praver 2018). Овакав начин учења ученицима омогућава два приступа која побољшавају исход учења, нарочито на нижим нивоима знања језика (Ashcroft, Cvitkovic, and Praver 2018) - свесно учење (Nation 2006; Hung 2015) и интервално учење нових речи (Ashcroft, Cvitkovic, and Praver 2018). Узимајући у обзир наведено, сматрамо да ћемо израдом модела представљеног у овом раду испитати неколико методолошких проблема, и допринети дигитализацији наставе језика на Рударско-геолошком факултету Универзитета у Београду.

Како се рад заснива на интердисциплинарном приступу, и циљеви рада су двојаки — педагошки и методолошки. Педагошки циљ је провера тренутног познавања геолошког и академског вокабулара студената и сарадника Рударско-геолошког факултета и унапређивање наставе учествовањем у развоју дигиталних наставних материјала. Методолошки циљ је да испитамо могућности примене LSA на процену одговора из домена геологије и академског вокабулара, и на форму дефиниције, као и да испитамо у којој мери модел корелира са проценама евалуатора. У складу са постављеним циљевима, полазним хипотезама претпостављамо да ће креирање система олакшати даљи процес дигитализације наставе, као и да ће примена LSA у оквиру дигиталних лексичких картица бити успешна.

## 2. Учење вокабулара коришћењем лексичких картица

Знање речи обухвата познавање њене форме, значења и употребе (Nation 2006). На учење вокабулара утичу и време које ученик проведе учећи, као и укљученост ученика у процес учења. Показало се да

при учењу вокабулара када је у питању време, најбоље резултате даје интервално учење (енгл. *spaced/distributed learning*), односно учење у више мањих сесија са паузама, где се дужина паузе између сваког интервала постепено повећава (Nation 2006). Када је у питању укљученост ученика у процес учења, према неким ауторима, имплицитно учење примењује се на учење форме речи и резултат је фреквенције изложености речима, док се значење речи, које је апстрактно и спада у домен семантике, усваја експлицитним учењем. При експлицитном учењу, користе се различита наставна средства, као што су речници, листе речи и лексичке картице (Nation 2006; Hung 2015; Ma 2009).

Неколико истраживања показало је да лексичке картице (енгл. *flash-cards*) дају најбоље резултате када је у питању усвајање вокабулара другог језика, нарочито на нижим нивоима познавања језика (Spiri 2008; Nakata 2008; Hung 2015; Averianova 2015; Yüksel, Mercanoğlu, and Yılmaz 2022). Лексичке картице истовремено омогућавају експлицитно и интервално учење вокабулара, те усвајање форме, значења и употребе речи у контексту (Ma 2009). Данас, ученицима су на располагању и дигиталне лексичке картице, за које се показало да имају одређене предности у односу на папирне — ученик одмах добија објашњење и исправку граматичких и правописних грешака, могуће је користити слике и звук, и може им се приступити са различитих уређаја. Такође, омогућавају учење било када и било где, и одличан су материјал за интервално учење. Могу се користити и за учење и за проверу знања, и омогућавају гејмификацију учења вокабулара. Неки од најчешће коришћених постојећих система су *Anki*[1] и *Quizlet*[2] (Ashcroft, Cvitkovic, and Praver 2018).

Учење вокабулара енглеског језика из домена геолошке терминологије помоћу дигиталних картица испитивано је у Beko, Obradović, and Stanković (2015). Неки од проблема истакнутих у раду су немогућност студената да пронађу адекватан метод за учење нових речи, низак ниво знања енглеског језика на почетку студија, као и мањак превода терминологије на српски језик, што додатно отежава задатке који захтевају превод. Како ће наш модел бити једнојезичан, нећемо понудити решење за последњи наведени проблем.

Постојећи системи који се користе у оквиру Рударско-геолошког факултета су тезаурус геолошке терминологије на српском и енглеском

1. Anki flashcards, приступљено 20. маја 2023.
2. Quizlet flashcards, приступљено 20. маја 2023.

језику који садржи око 2800 речи, као и дигитални ресурс рударске терминологије *RudOnto*[3] (Beko, Obradović, and Stanković 2015). Развијен је и систем дигиталних картица *RGF Flashcards*, коришћењем програма *Anki*, које су потом интегрисане у *Moodle* платформу.[4]

С обзиром на то да ће се дигиталне лексичке картице користити у настави која прати уџбеник заснован на наставном методу CLLIL, која спаја учење садржаја из одређене стручне области са учењем језика (Beko 2013; Đerić 2019; Baten, Van Hiel, and De Cuypere 2020), при чему се од ученика очекује познавање енглеског језика на нивоу C1, картице могу помоћи ученицима који ступају на факултет са нижим нивоом познавања енглеског језика да брже и лакше усвоје вокабулар и самим тим лакше прате наставу и наставне материјале.

## 3. Латентна семантичка анализа LSA

Латентна семантичка анализа (енгл. *Latent Semantic Analysis*, LSA) је теорија и метода за екстракцију и представљање значења речи у контексту, статистичким израчунавањима примењеним на велики корпус текста (Landauer et al. 1997). Досадашња истраживања показала су да LSA у великој мери на добар начин представља когнитивне способности као што су усвајање вокабулара, категоризација речи, семантичко примовање, разумевање дискурса, и процена ваљаности есеја (Landauer, Foltz, and Darrell 1998).

Приликом израде LSA модела на почетку правимо матрицу докумената и термина, чиме се формира семантички простор саздан од свих термина и докумената у корпусу (Deerwester et al. 1990; Landauer, Foltz, and Darrell 1998; Lemaire and Dessus 2003). У наредном кораку, функција тежине примењује се на сваку ћелију у матрици, при чему се мале тежине додељују високофреквентним терминима, а велике тежине терминима који се појављују у неким, али не свим документима (Deerwester et al. 1990; Martin and Berry 2013). Након овога, врши се декомпозиција матрице на сингуларне вредности (енгл. *Singular Value Decomposition*, SVD). SVD (формула 1) је кључни део LSA, јер омогућава представљање значења речи у односу на контекст у којем се јављају (Lemaire and Dessus 2003). Нека су $m$ и $n$ природни бројеви и нека је $M$ произвољна матрица $m \times n$. Тада постоји декомпозиција матрице $M_{mn}$:

3. RudOnto тезаурус, приступљено 20. маја 2023.
4. Moodle, приступљено 20. маја 2023.

$$M = U \times S \times V^T \tag{1}$$

где је:

- $M$: ортогонална $m \times n$ матрица (матрица докумената и термина), где $m$ представља број докумената, док $n$ представља број термина;
- $U$: ортогонална $m \times r$ матрица докумената и тема, где $m$ представља број докумената, док $r$ представља број тема;
- $S$: дијагонална $r \times r$ матрица, чије су све вредности осим на дијагонали једнаке 0. Вредности у $S$ представљају колико свака латентна тема објашњава варијансу у подацима;
- $V$: ортогонална $n \times r$ матрица термина и тема, где $n$ представља број докумената, док $r$ представља број тема.

Множењем трију матрица добијених у SVD можемо приближно реконструисати оригиналну матрицу $M$. Димензионалност редуковане матрице $M_k$ треба да буде оптимална, и да тачно осликава односе између елемената у оригиналној матрици $M$ (Landauer, Foltz, and Darrell 1998). За проверу ваљаности броја димензија, узима се спољни критеријум валидације (Landauer, Foltz, and Darrell 1998). У нашем случају, ово ће бити мера косинусне сличности између одговора студента и тачног одговора (Rahutomo, Kitasuka, and Aritsugi 2012).

До сада, LSA је коришћена у многим паметним играма у сврхе процене тачности одговора, давања повратне информације о одговору, одговарање на упите студената, процену тачности и кохерентности есеја и учење вокабулара. Приликом процене тачности есеја показала је висок степен корелације са проценама евалуатора (Landauer et al. 1997; Graesser et al. 2000; Lemaire and Dessus 2003; Landauer, Laham, and Foltz 2003; Dikli 2006; Lafourcade and Zampa 2009; Picca, Jaccard, and Eberlé 2015).

## 4. Опис и припрема улазних података

Улазни подаци састоје се од три дела. Први део улазних података су текстови написани за уџбеник у припреми проф. др Лидије Беко - 12 лекција са по три текста о одређеној теми. Други део улазних података је вокабулар који прати сваку лекцију. Вокабулар укупно има 962 речи, а подељен је у три категорије - општи вокабулар (663 речи), геолошки вокабулар (280 речи) и минерали (18 речи). Трећи

део података су одговори испитаника прикупљени на основу тестова окачених у оквиру курсева Енглески 1–4 на платформи *Moodle* Рударско-геолошког факултета.

У сврхе прикупљања одговора направљене су три групе теста за три групе испитаника, од којих је свака имала исте задатке са различитим примерима. Све групе теста формиране су прилагођавањем задатака коришћених у Jhean-Larose et al. (2010), а шести задатак прилагођен је у односу на оригинални тест. Описи задатака са примерима налазе се у табели 1.

| **Бр** | **Задатак** | **Пример** |
|---|---|---|
| 1 | оценити десет дефиниција као тачне TRUE или нетачне FALSE | *fossilisation is a process in which parts of a dead animal or plant being turned into a part of sediment and thereby preserved* TRUE |
| 2 | упарити 11 речи са њиховим дефиницијама | *fern – а) a vascular plant with complex fronds and sporangia on the leaves' surface where asexual spores are found* |
| 3 | направити дефиницију речи и у оквиру дефиниције искористити задате појмове | помоћу појмова - *collection, fragment* написати дефиницију речи *debris (a collection of fragments of rocks)* |
| 4 | спојити делове дефиниција А и Б, а затим их повезати са одговарајућим речима | Part А: *pertaining to complex protoplasmic life-forms*, Part В: *with a vesicular nucleus and various cytoplasmic organelles - eukaryotic* |
| 5 | скуп задатих особина неког појма оценити као тачан TRUE или нетачан FALSE | *embed is... а) to be placed* TRUE; *б) within something* TRUE; *в) so that the part can be easily removed* FALSE |
| 6 | објаснити **зашто** долази до одређене појаве и **шта** је одређена појава | *Explain what global warming is and why it happens; What is seafloor?* |
| 7 | објаснити **како** долази до нечега | *How are sedimentary rocks formed?* |
| 8 | написати дефиниције за 10 речи | *mineral (a naturally occurring inorganic substance with a characteristic chemical composition)* |

Табела 1: Тест помоћу којег су прикупљани подаци

Овако састављени тестови имплементирани су на *Moodle* платформи која је додатно опремљена проширењем H5P[5] и подељени са испитаницима. Тест је попунило 14 испитаника, а прикупљен је 451 одговор. Како би наши испитаници остали анонимни, сваком испитанику додељен је јединствени ID. Највише одговора прикупљено је за прву, а најмање за трећу групу теста. Евалуација одговора вршена је у два корака, прво је сваком одговору додељена оцена на скали од 1 до 5 (табела 2).

| **Оцена** | **Значење** |
|---|---|
| 1 | одговор је сасвим нетачан |
| 2 | део одговора је тачан |
| 3 | одговору недостаје део дефиниције |
| 4 | одговор је скоро сасвим тачан |
| 5 | одговор је потпун |

Табела 2: Скала процене евалуатора

Одговоре којима је додељена оцена 1, означили смо као нетачне (N), а одговоре са осталим оценама као тачне (T). Критеријум за процену тачности одговора испитаника био је поређење са златним стандардом – дефиницијом одреднице из уџбеника, и знање евалуатора. При процени тачности нисмо узимали у обзир граматичку и правописну исправност одговора. Услед недоследног излазног резултата петог задатка, овај задатак избачен је из анализе. За анализу помоћу LSA узети су 3, 6, 7. и 8. задатак. Овим смо добили 238 одговора, а након уклањања питања на која испитаници нису одговорили, остало нам је 72 одговора за евалуацију.

### 4.1 Обрада текста

Припрема текста рађена је у складу са методама изнађеним у литератури (Deerwester et al. 1990; Dikli 2006; Lifchitz, Jhean-Larose, and Denhière 2009), које смо прилагодили нашим циљевима и подацима. Текст је прво лематизован коришћењем библиотеке *SpaCy*,[6] а када

5. H5P проширење, приступљено 22. маја 2023.

6. SpaCy библиотека, приступљено 22. маја 2023.

| **Оригинални текст** | **Обрађен текст** |
|---|---|
| Most people today are familiar with mineral water and the perennial debate, as to whether still or sparkling is better. | most people today be familiar with mineral water and the perennial debate as to whether still or sparkle be well |
| Groundwater stored in subterranean aquifers has always been extracted for human use through the digging of wells. | groundwater store in subterranean aquifer have always be extract for human use through the digging of well |
| The conversion of sediment to rock is known as lithification transformation or diagenesis, and tends to involve two stages – compaction and cementation. | the conversion of sediment to rock be know as lithification transformation or diagenesis and tend to involve two stage compaction and cementation |

Табела 3: Обрађен текст

смо добили лематизоване сурогате за сваки део података, из њих смо уклонили интерпункцију и специјалне карактере коришћењем регуларних израза и текст пребацили у мала слова (енгл. *lower case*). Из одредница смо уклонили и латинске множине речи (*data sing. datum* и *hypothesis pl. hypotheses*). Пример обрађеног текста приказан је у табели 3. Реченице су преузете из различитих текстова.

У дефиницијама, глаголи нису доследно лематизовани, нпр. "an act of splitting into category" реч *splitting* није лематизована у *split*. Прегледом других примера са сличном структуром (глагол-предлог-глагол или именица-глагол-предлог), нпр. *an action of take something, the process of break something down* нисмо уочили исту грешку. Такође, глагол *unstratified* је у одговору где се налази самостално лемазитован у *unstratifie*, док је у одговору *incoherent loose unstratified* остао у истом облику. Латинске речи, нпр. *antennae* или *pinaceae*, нису лематизоване.

## 5. Израда модела за процену тачности одговора

За израду LSA користили смо библиотеку *Scikit-Learn*.[7] Прва израђена матрица је TF-IDF (енгл. *term frequency–inverse document*

7. Scikit-Learn библитека, приступљено 22. маја 2023.

*frequency*, TF-IDF) матрица, где се у редовима налазе документи, у колонама термини, а у ћелијама релативне фреквенције термина у сваком од докумената (Jurafsky and Martin 2023).

Приликом израде матрице потребно је одлучити са колико термина ће сваки документ бити описан. Испробавањем различитих опција између 700 и 5000 термина, одлучили смо се да број термина у TF-IDF матрици текстова лекција буде 1.000, да минимална фреквенција буде три и да се одбацују термини који се појављују у више од 80% докумената. Уклоњене су и стоп речи, које су комбинација стоп речи за енглески језик из NLTK библиотеке[8] и стоп речи специфичних за наше текстове (*km, kmh, mm, metre, one, two, three, etc, yet, well...*). Како применом ових параметара на дефиниције и одговоре нисмо добили добре резултате, одабрали смо да TF-IDF матрица овде има 700 димензија, а да минимална и максимална фреквенција буду 1 и 100% докумената. Стоп речи уклоњене из одговора и дефиниција су само неодређени и одређени члан — *a/an, the*.

Параметри SVD исти су за све делове података. Како бисмо утврдили да ли смо добро одредили број тема, за сваку добијену тему издвојили смо 15 термина са највећим тежинама, затим смо проверили да ли између тема постоје превелика преклапања, или да ли модел нешто што сматрамо да је требало, није издвојио. Овиме смо одлучили да број тема буде 10. Прегледом првих 100 термина, доделили смо називе темама. Неке теме, *Topic0, Topic1, Topic4*, садрже општије термине, који се провлаче кроз већину текстова. С друге стране, у темама *Topic3, Topic5* и *Topic7* се налазе термини који се односе на специфичне области геологије, као што су тектонске плоче, вулканологија и ерозија (табела 4).

Након што смо спремили векторе за све делове података, измерили смо косинусне сличности између свих текстова како бисмо пронашли међусобно најсличније, а затим смо израчунали скор сваког одговора мерењем косинусне сличности одговора и: а) вектора текстова лекције у којој се реч коју испитаник дефинише налази; б) вектора тачног одговора (златног стандарда); в) вектора њему најсличнијег одговора. Што је скор сличности између вектора документа $A$ и вектора документа $B$ већи, већа је и повезаност између докумената (Rahutomo, Kitasuka, and Aritsugi 2012). На основу добијеног скора одговоре смо класификовали према тачности коришћењем алгоритма K најближих суседа (енгл. K-*Nearest Neighbor*, KNN) (Li, Yu, and Lu 2003; Peterson 2009), прво у две

8. NLTK библитека, приступљено 22. маја 2023.

| **Бр. теме** | **Назив теме** | **Термини са највећим тежинама** |
|---|---|---|
| Topic0 | Earth Formation | mineral, cycle, earth, deposit, flow, sedimentary, igneous, material, soil, metamorphic, sediment, begin, grain, metamorphism, plant |
| Topic1 | Minerals | mineral, metamorphism, grain, metamorphic, igneous, metamorphic rock, pressure, crystal, magma, ore, deposit, chemical, metallic, thermal, colour |
| Topic2 | Erosion | flow, soil, particle, stream, slope, erosion, debris, landslide, glacial, material, groundwater, sand, glacier, velocity, move |
| Topic3 | Tectonic Plates | plate, earthquake, wave, cycle, tectonic, magma, continental, oceanic, magnetic, magnetic field, earth, activity, stress, volcano, temperature |
| Topic4 | Rock Formation | sedimentary, cycle, sediment, metamorphic, igneous, sedimentary rock, strata metamorphic rock, metamorphism, erosion, grain, plate, rock cycle, igneous rock, pressure |
| Topic5 | Volcanology | magma, lava, grain, volcano, slope, eruption, volcanic, viscosity, period, landslide, hazard, volcanic eruption, debris, era, mesozoic |
| Topic6 | Weathering | wave, earthquake, magnetic, date, particle, magnetic field, metamorphism, stress, erosion, sediment, grain, field, age, sedimentary, strata |
| Topic7 | Landslides | slope, landslide, soil, debris, hazard, cycle, trigger, activity, fall, downslope, mitigation, metamorphism, metamorphic rock, metamorphic, angle |
| Topic8 | Dating | earth, strata, magma, date, age, eruption, lava, idea, satellite, remote, atom, history, feature, geological, sedimentary |
| Topic9 | Fossils | oil, wave, earthquake, coal, trap, organic, sedimentary, sedimentary rock, weathering, carbon, plant, oil gas, soil, gas, type |

Табела 4: Издвојене теме из текстова и термини са највећим тежинама за сваку тему

категорије — тачно (т) и нетачно (н), а затим у више категорија према критеријумима описаним у табели 2.

### 5.1 Расподела текстова, дефиниција и одговора по темама

Пре него што класификујемо одговоре, погледаћемо расподелу података по темама, како бисмо видели колико добро наш модел ради. Расподела вредности по темама најнеуједначенија је код текстова, док је код дефиниција и одговора нешто уједначенија, али и даље неправилно распоређена. Претпостављамо да је разлог мањој стандардној девијацији (енгл. *standard deviation*, STD) (Urdan 2005) података у дефиницијама и одговора кохерентнија форма текста у односну на текстове.

У текстовима, максималне вредности тема крећу се од 0,5617 код теме *Volcanology*, па до 0,3638 код теме *Dating*, док се минималне вредности крећу између 0,3878 за *EarthFormation*, па све до -0,001 за тему *Dating*, а вредности унутар самих тема распршене су у односу на просек. Код дефиниција, максималне вредности тема нешто су уједначеније. Највећу максималну вредности има тема *Weathering* (0,7067), а најмању *Landslides* (0,3547). Готово све минималне вредности тема су негативне, изузев теме *EarthFormation* са минималном вредношћу од 0,0193. Док су неке дефиниције геолошких појмова добро распоређене по темама, на пример *debris*[9] има високе вредности у темама *EarthFormation*, *Weathering* и *Landslides*, дефиниције речи *fossil*,[10] *fossilised*, *fossilisation* немају високе вредности у теми *Fossils*, што значи да према нашем моделу, ова тема не описује ове појмове у довољној мери. Расподелу академских речи по темама нешто је теже евалуирати како се теме односе на геолошке области. Фразални глагол *wear away*[11] има високе вредности код тема *EarthFormation*, *Erosion* и *Weathering*, што се уклапа са његовим значењем. Ипак, речи општег вокабулара као што су прилози *therefore*, *yet*, *anyway*[12] итд, вероватно је било најпроблематичније распоредити по темама, јер се понављају кроз све теме.

Максималне вредности тема у одговорима испитаника варирају, од највеће 0,7042 код теме *TectonicPlates*, па до најмање максималне вредности од свега 0,3612 за тему *Landslides*. Минималне вредности углавном су негативне, и крећу се од -0,5557 код теме *Minerals*, па до 0 за вредност теме *EarthFormation*. Одговори на исто питање углавном имају сличну расподелу вредности по темама па се тако на пример

9. the remainders of something destroyed
10. parts of animals or plants that have been hardened and preserved in sediment
11. an action of gradually eroding or grinding something down
12. стога, ипак, свакако

код одговора на питање *global warming*[13] највеће вредности у већини одговора јављају код теме *EarthFormation*, а најмање вредности су код тема *Erosion* и *Landslides*. Код кратких, непотпуних одговора, вредности свих тема су 0.

## 6. Резултати

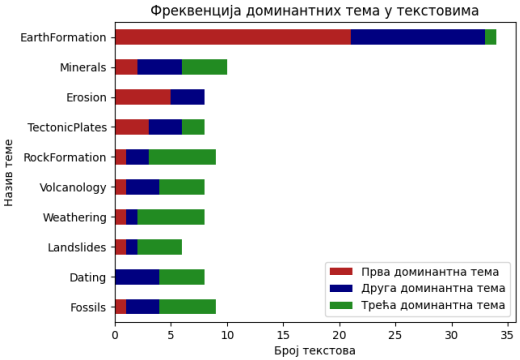
Излагање резултата започећемо представљањем доминантних тема у текстовима, дефиницијама и одговорима испитаника. У наредном одељку, представићемо начин израчунавања тачности одговора и урадити анализу добијених резултата. На крају, погледаћемо резултате добијене класификацијом одговора према тачности коришћењем KNN алгоритма.

### 6.1 Доминантне теме у текстовима, дефиницијама и одговорима испитаника

За сваки узорак у сваком делу података издвојили смо три доминантне теме. У текстовима лекција тема *EarthFormation* најфреквентнија је и као прва и као друга доминантна тема. Као прва доминантна тема јавља се у 21 документу, као друга у 11 докумената, а као трећа у два (графикон на слици 1). Једина тема која се ниједном не јавља као прва доминантна је тема *Dating*, али је на другом месту по фреквенцији као друга доминантна тема, заједно са темом *Minerals*. Све теме се јављају на месту друге доминантне теме, од чега се тема *Landslides* јавља само једном, у тексту *Mass Wasting and Types of Landslides*. Једина тема која се ниједном не појављује као трећа доминантна је тема *Erosion*.

Када су у питању вредности доминантних тема, најмању распршеност видимо код прве доминантне теме, која се креће од око 0,3 до око 0,6, док су вредности друге и треће доминантне теме нешто распршеније и ниже, испод 0,1 до око 0,4 за другу, а за трећу се крећу од негативних, па све до око 0,4 (слика 2). Доминантне теме текстова углавном су добро одређене, са изузетком текста *Mineral Evolution*, код којег се тема *Minerals* не јавља међу доминантним, док се јавља у наредном тексту исте лекције, *Physical Property*, која говори о

13. an increase in global temperature due to various factors such as an increase in carbon dioxide emission and pollution with a potentially catastrophic outcome



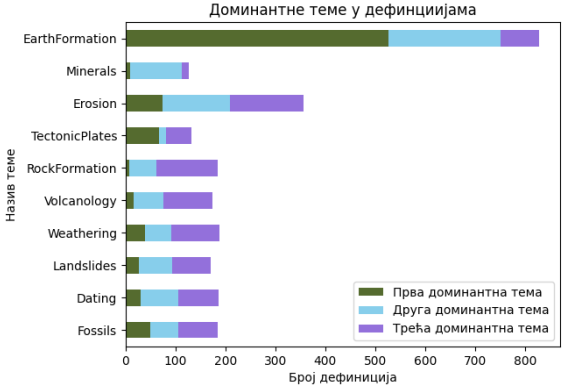

Слика 1: Фреквенција доминантних тема у текстовима лекција.

настајању и физичким особинама минерала. Такође, у тексту *Weathering* истоимена тема не налази се међу доминантним.


Слика 2: Вредности доминантних тема у текстовима лекција.

На основу добијених резултата можемо закључити да постоје велике разлике у вредностима прве и друге доминантне теме у документу. Ипак, друга и трећа доминантна тема често нам специфичније говоре о чему се ради у тексту, па приликом анализе све три доминантне теме треба узети у обзир.

Као и код текстова, тема *EarthFormation* је и у дефиницијама најфреквентнија као прва и друга доминантна тема. За њом следи тема *TectonicPlates*, а блиске су јој и *Fossils, Volcanology* и *Landslides*. Тема *Minerals* се најмање пута јавља као прва доминантна, али је као друга најчешћа (слика 3).




Слика 3: Доминантне теме у дефиницијама.

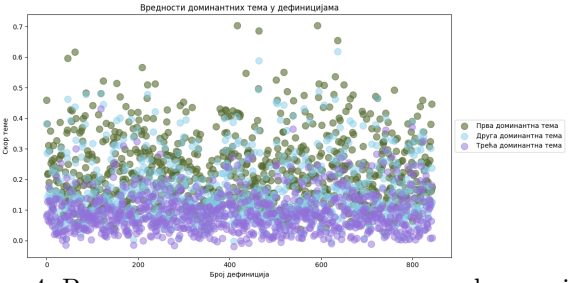
У дефиницији речи *glaciation*,[14] као доминантне теме додељене су *Weathering, EarthFormation* и *Volcanology*. Са друге стране, дефиниција речи *earthflow*[15] смештена је првенствено у теме *Fossils, EarthFormation, TectonicPlates*, иако бисмо је интуитивно пре сврстали у *Erosion, Landslides* и *Weathering*. Када су у питању вредности доминантних тема, највеће разлике видимо у вредностима прве доминантне теме, која се у дефиницијама креће од преко 0,7, па до испод 0,1. Вредности друге доминантне теме нешто су ниже од прве, али подједнако расејање, док су вредности треће доминантне теме углавном релативно ниске, испод 0,4, и релативно хомогене (слика 4).

У одговорима испитаника тема *EarthFormation* такође је најфреквентнија на сва три места, при чему се као прва доминантна тема јавља у 31, као дуга у 25, као трећа у 18 одговора. Као прва доминантна тема на другом месту налази се *Landslides* са највећим вредностима у 10 одговора, а као друга *Erosion* у 12 одговора. Фреквенције тема на месту треће доминантне су поприлично уједначене, изузев теме *Landslides* која се јавља свега два пута (слика 5).

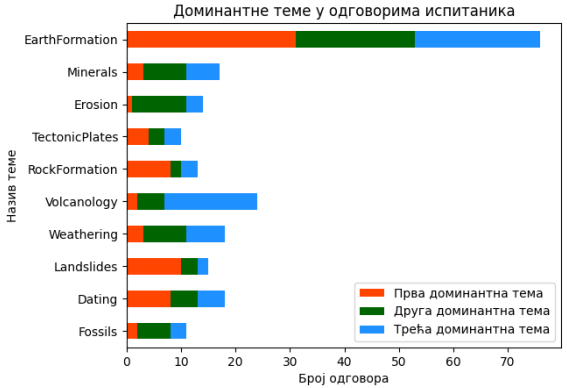
Дужим одговорима на исто питање углавном су додељене исте доминантне теме. Тако су у свим одговорима на питање *global warming* доминантне теме *EarthFormation, Volcanology* и *Minerals* или *Dating*.

14. used for referring to geological processes of a glacier – its formation, movement, and recession

15. a downslope movement of unconsolidated material, usually caused by percolation of water between the loose particles




Слика 4: Вредности доминантих тема у дефиницијама.




Слика 5: Доминантне теме у одговорима испитаника.

Код одговора на питање *sedimentary rock*[16] као прве две доминантне теме понављају се *EarthFormation* и *Landslides*, док се као трећа смењују *Fossils, Dating* и *Erosion*. Нулте вредности тема објашњене су нераспоређеним одговорима (слика 6).

Након што смо погледали расподелу тема, њихових вредности, те доминантних тема у свим деловима података, упоредили смо издвојене доминантне теме за исте појмове у одговорима, дефиницијама и текстовима лекција где се појам јавља (табеле 5 и 6).

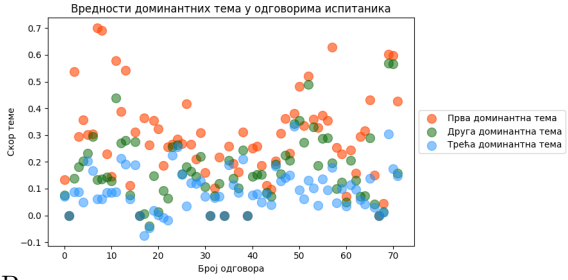
На пример, код речи *transpire* издвојене доминантне теме у одговору, дефиницији и текстовима се не покапају, а вредности доминантних тема разликују се у свим деловима података (табела 5). Код појма *global warming* уочили смо велики степен преклапања доминантних тема у

16. a type of rock formed by accumulation and cementation of transported sediments by means of water, wind, glacier, or gravity

| Термин | Дефиниција | Извор | Вредност теме | | | Назив теме | | |
|---|---|---|---|---|---|---|---|---|
| | | | I | II | III | I | II | III |
| transpire | transpire an action of discharge liquid through a plant discharge perspire excrete flow | дефиниција | 0,2037 | 0,1603 | 0,0302 | Erosion | Earth Formation | Fossils |
| transpire | release liquid through opening in leave of a plant | одговор | 0,1478 | 0,0435 | 0,0418 | Earth Formation | Weathering | Volcanology |

| Лекција | Текст | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| metamorphic rocks and gemstones | the causes and definition of metamorphism | текст | 0,4378 | 0,4111 | 0,2329 | Earth Formation | Minerals | Rock Formation |
| | metamorphic textures | | 0,5223 | 0,4447 | 0,1867 | Minerals | Earth Formation | Volcanology |
| | gemstones associated with metamorphism | | 0,5269 | 0,2629 | 0,0683 | Earth Formation | Minerals | Rock Formation |

Табела 5: Доминантне теме за термин *transpire*




Слика 6: Вредности доминантних тема у одговорима испитаника.

одговорима, дефиницији и текстовима (табела 6). Претпостављамо да је разлог великог преклапања то што на ово питање имамо више дужих одговора који садрже термине који имају велике тежине у темама, па је тиме олакшано и распоређивање.

### 6.2 Мерење сличности између текстова

Коришћењем косинусне сличности, измерили смо сличности између вектора добијених након примене SVD на TF-IDF матрицу свих текстова у уџбенику, а затим пронашли међусобно најсличније текстове. На основу овога, добили смо увид у то колико је добро наш LSA модел пронашао латентне теме у текстовима.

Анализом добијених резултата можемо закључити да су латентне теме у текстовима добро одређене, и да међусобно најсличнији текстови говоре о истим темама. Тако је на пример у тексту о Вагнеровој хипотези, у којем се објашњава како ова теза претпоставља постојање Пангее, најсличнији текст о тектонским плочама, а тексту о вулканима најближи је текст о вулканском камењу (табела 7).

### 6.3 Израчунавање тачности одговора

У овом делу рада, представићемо резултате добијене мерењем косинусне сличности вектора одговора A са: (1) вектором текстова лекције у којој се реч коју је испитаник требало да дефинише налази; (2) вектором тачног одговор; (3) њему најсличнијим одговором Б. Како бисмо видели да ли постоје разлике у скору тачних и нетачних одговора, погледаћемо расподелу вредности коначног скора одговора по тачности у две и више категорија на основу процена евалуатора.

| **Термин** | **Дефиниција** | **Извор** | **Вредност теме** | | | **Назив теме** | | |
|---|---|---|---|---|---|---|---|---|
| | | | **I** | **II** | **III** | **I** | **II** | **III** |
| global warming | global warming an increase in global temperature due to various factor such as increase carbon dioxide emission and pollution with a potentially catastrophic outcome | дефиниција | 0,1256 | 0,1159 | -0,0016 | Volcanology | Earth Formation | Dating |
| global warming | rise in global temperature of the earth due to carbon dioxide emission | одговор | 0,3257 | 0,2093 | 0,0006 | Earth Formation | Volcanology | Minerals |
| **Лекција** | **Текст** | | | | | | | |
| fossils through times | fossil formation and palynology | текст | 0,4997 | 0,1832 | 0,0836 | Earth Formation | Fossils | Volcanology |
| | palaeozoic era | | 0,3641 | 0,2228 | 0,0228 | Earth Formation | Minerals | Rock Formation |
| | mesozoic era and cenozoic | | 0,3702 | 0,2567 | 0,0263 | Earth Formation | Volcanology | Weathering |

Табела 6: Доминантне теме за термин *global warming*

| **Наслов текста а** | **Наслов текста б** | **Сличност** |
|---|---|---|
| palaeozoic era | mesozoic era and cenozoic | 0,9776 |
| wegener s hypothesis, seafloor spreading, convection cells | tectonic plates | 0,8980 |
| volcanoes | igneous rocks | 0,8229 |
| the causes and definition of metamorphism | metamorphic textures | 0,9606 |
| coal as a fossil fuel | oil and natural gas mineral oil | 0,7726 |

Табела 7: Примери међусобно најсличнијих текстова

При поређењу одговора и текстова, поредили смо одговоре испитаника са три текста лекције у којој се дефинисана реч јављa, а затим израчунали аритметичку средину добијених косинусних сличности текстова како бисмо добили сличност са целом лекцијом. Добијене резултате треба узети са задршком, јер се вектори појединачних текстова засигурно разликују од вектора целе лекције. Код дужих одговора уочене су веће сличности са текстовима, нарочито када су у питању геолошки појмови као што су *global warming*, *sedimentary rock*. Код одговора на питање *hypothesis*[17] сличност је у једном одговору 0, док је у другом поприлично висока (табела 8).

Као што видимо, расподела вредности сличности тачних одговора и текста нешто су више него код нетачних одговора. Код расподеле вредности у пет категорија тачности одговора, највећи распон имамо у вредностима категорија 3 и 5. Најниже вредности имају одговори са оценом 1. Аутлејери (енгл. *outliers*) су одговори на питања *chemical weathering*[18] и *convergent*[19] (слика 7).

Косинусна сличност одговора и тачног одговора вероватно је најбољи показатељ тачности и потпуности одговора испитаника. Сваки одговор поређен је само са једном дефиницијом, тако да у овом кораку поредимо

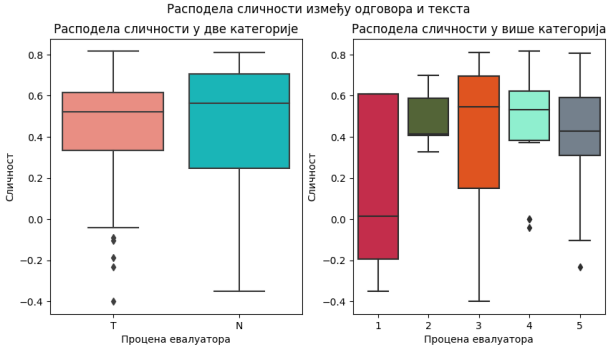
17. an assumption resting on previous knowledge, but has not yet been proven true or false

18. transformation of rocks through chemical reactions when exposed to air or water containing dissolved elements

19. used for describing two or more objects or ideas moving towards the same point or developing in the same direction

| **идп** | **иди** | **Питање** | **Измерене косинусне сличности** | | | | **T/N** | **1–5** |
|---|---|---|---|---|---|---|---|---|
| | | | **Текст** | **Дефин.** | **Одговор** | **Скор** | | |
| 6 | 3 | global warming | -0,2713 | 0,7325 | 0,8629 | 0,8004 | T | 3 |
| 6 | 6 | global warming | 0,7591 | 0,8259 | 0,9680 | 0,8998 | T | 4 |
| 7 | 6 | sedimentary rock | 0,5107 | 0,6346 | 0,9422 | 0,8033 | T | 5 |
| 7 | 1 | sedimentary rock | -0,2183 | 0,5181 | 0,9317 | 0,7538 | T | 5 |
| 3 | 4 | hypothesis | 0,0000 | -0,0094 | 0,7816 | 0,5527 | T | 5 |
| 3 | 3 | hypothesis | 0,7104 | 0,7017 | 0,8666 | 0,7885 | T | 4 |

Табела 8: Резултати измерених косинусних сличности —- примери за сличност одговора и текста; ИДП – ID питања, ИДИ – ID испитаника




Слика 7: Расподела сличности између одговора испитаника и текстова.

сличност два кратка текста, па добијамо увид у то како наш LSA модел заправо ради на кратком тексту и на форми дефиниције. На пример, у питању *convergent* испитаник 4 дефинисао је реч као математички појам, што није нетачно, међутим, није значење речи које се тражи, па је и сличност са тачним одговором поприлично ниска (табела 9). Такође, дефиниција речи *straightforward*,[20] где имамо недовршен одговор, има веома ниску сличност са тачним одговором (табела 9). Велики број одговора који су тачни и потпуни има велике сличности са тачним одговором. Са друге стране, неки тачни одговори имају негативне

20. occurring without obstacles or irregularities, in a simple manner

косинусне сличности са тачним одговором (*petrologist*)[21] (табела 9). Претпостављамо да је разлог овоме дужина одговора.

| **идп** | **иди** | **Питање** | **Измерене косинусне сличности** | | | | **T/N** | **1–5** |
|---|---|---|---|---|---|---|---|---|
| | | | **Текст** | **Дефин.** | **Одговор** | **Скор** | | |
| 3 | 3 | convergent | 0,0000 | 0,2885 | 0,9992 | 0,7354 | T | 4 |
| 3 | 4 | convergent | -0,3299 | 0,2999 | 0,9992 | 0,7377 | N | 1 |
| 8 | 3 | straightforward | -0,0291 | 0,0011 | 0,9627 | 0,6807 | N | 1 |
| 8 | 6 | petrologist | 0,1837 | 0,2238 | 0,9803 | 0,7110 | T | 3 |

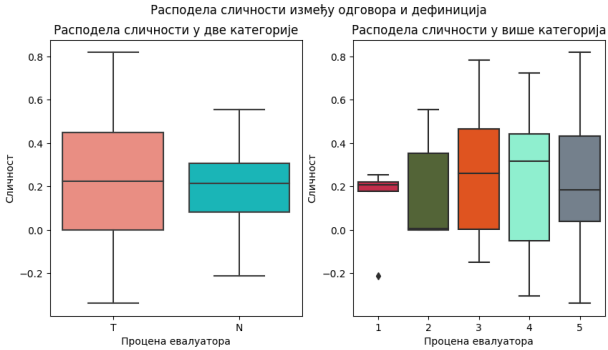
Табела 9: Резултати измерених косинусних сличности —- примери за сличност одговора и тачног одговора; ИДП - ID питања, ИДИ - ID испитаника

Такође, на резултате је могла утицати и велика тежина функционалних речи (*be*, *of*, *in*, *to*, *or*, *by* итд.) у темама дефиниција. Уклањање функционалних речи са великим тежинама из дефиниција могло би решити овај проблем.

Као што можемо видети, код тачних одговора имамо већи распон вредности него код нетачних. Ниске вредности код тачних одговора могу постојати услед кратких одговора који су процењени као тачни (слика 8). У сличностима одговора испитаника и тачног одговора, највећи распон у вредностима уочавамо код одговора оцењених са као потпуно тачних (оцена 5). Ниске сличности одговора у овој категорији и тачног одговора уочили смо код одговора где је испитаник користио речи синонимне онима у тачном одговору (слика 8).

Трећа измерена сличност је сличност свих одговора, коју смо измерили на исти начин као сличност између свих текстова. За разлику од текстова, најсличнији одговори нису најбоље одређени. Добијене сличности крећу се од око 0,7 до око 0,9. У случајевима где је вектор одговора 0, добили смо највећу сличност 0 са првим одговором у подацима (табела 10). Као најсличнији одређени су одговори који не деле заједничке термине, затим одговори на исто питање који имају доста заједничких термина, али и одговори на различита питања који садрже

21. a scientist in the field of petrology, studying rocks and how they are formed




Слика 8: Расподела сличности између одговора и дефиниција.

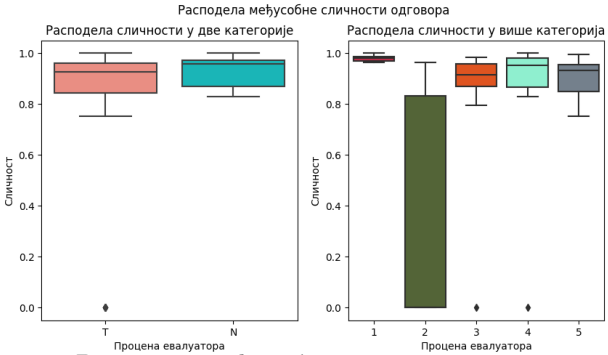
исте термине, нпр. *hydrological cycle*[22] и *seabed*,[23] где оба одговора садрже речи као што су *earth*, *ocean*, *surface* (табела 10).

Расподела између тачних и нетачних одговора је готово иста, пошто су узимане вредности највећих косинусних сличности. У вредностима међусобне сличности одговора, највећи распон вредности имамо код одговора са оценом 2 (слика 9).

| **Питање а** | **Одговор а** | **Питање б** | **Одговор б** | **Сличност** |
|---|---|---|---|---|
| hydrological cycle | the hydrological cycle of the earth be the sum total of all process in which water move from the land and ocean surface to the atmosphere and back in form of precipitation | seabed | the seabed be the bottom of the ocean or the top surface of the earth in sea and ocean | 0,8685 |
| unconsolidate | unstratifie | backlash | adverse reaction to a recent development | 0,0000 |
| urbanisation | make an area more urban | urbanisation | be the process of make an area more urban | 0,9840 |

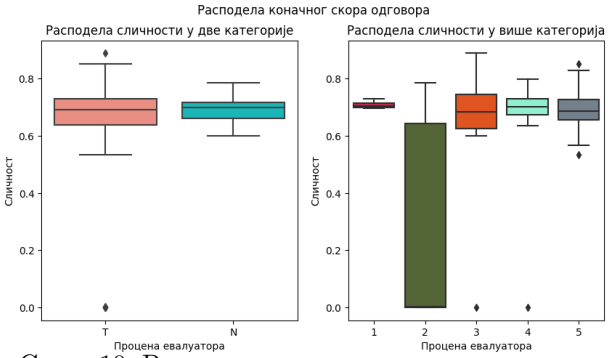
Табела 10: Међусобно најсличнији одговори

22. the representation of a continuous, circular movement of water through the atmosphere, where the physical state of water alters as it flows through the cycle
23. land at the bottom of the ocean




Слика 9: Расподела међусобне сличности свих одговора.

На крају, израчунали смо коначан скор сваког одговора. Најмања вредност у коначном скору је 0, и то код претходно поменутих кратких одговора. У дужим одговорима све три сличности су приближно једнаке. Код расподеле вредности коначног скора у двема категоријама, поново добијамо ниже вредности код тачних него код нетачних одговора, али и већи распон вредности тачних одговора. На високе вредности нетачних одговора вероватно је утицала сличност са најсличнијим одговором (слика 10). Код расподеле скора према пет категорија, вредности су најраспршеније код оцене 2, а најзбијеније код оцене 1, док више оцене имају релативно сличне коначне скорове (слика 10).




Слика 10: Расподела вредности скора одговора.

Напослетку можемо закључити да је LSA дала добре резултате када су у питању текстови лекција. Са друге стране, када су у питању дефиниције и одговори, резултати су нешто лошији, нарочито код проналажења скривених тема у веома кратким одговорима, те у одговорима у којима је испитаник требало да дефинише реч која не припада домену геологије. Такође, косинусне сличности тачних и нетачних одговора се веома мало разликују, па се доводи у питање овакав начин процене тачности.

### 6.4 Класификација одговора према тачности

На крају, извршили смо класификацију одговора према тачности у две категорије (енгл. *binary classification*) и у више категорија (енгл. *multiclass/multinominal classification*). За класификацију је коришћен KNN алгоритам (Li, Yu, and Lu 2003; Peterson 2009; Chen 2018), ознаке су оцене које је одговорима доделио евалуатор, а критеријум је коначни скор одговора. За евалуацију модела узети су одзив, прецизност и $F_1$ мера (Géron 2022).

а) класификација у две категорије

| | **N** | **T** |
|---|---|---|
| **N** | 0 | 4 |
| **T** | 0 | 11 |

б) класификација у више категорија

| | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **1** | 0 | 0 | 1 | 0 | 1 |
| **2** | 0 | 1 | 0 | 0 | 0 |
| **3** | 0 | 0 | 2 | 1 | 0 |
| **4** | 0 | 0 | 2 | 1 | 0 |
| **5** | 1 | 0 | 1 | 1 | 3 |

Табела 11: Матрице конфузије

При бинарној класификацији одговори су класификовани на тачне и нетачне. У подацима имамо 60 тачних и 12 нетачних одговора. Услед много већег броја тачних одговора у подацима, модел је све одговоре класификовао као тачне (табела 11а). Процењена прецизност модела је 73%, одзив 100%, а $F_1 = 0,85$. Пошто радимо са релативно малим скупом података од 72 узорка, па се самим тим тест скуп података састоји од

свега 15 узорака, ови резултати не дају реалну слику ваљаности модела, с обзиром на то да се подаци на тренинг и тест скуп деле насумично, и може се десити да сви узорци у тест скупу података имају исту ознаку.

Приликом класификације у више категорија, такође немамо једнаку фреквенцију категорија, где су најмање заступљени нетачни (5) и сасвим тачни одговори (7), па очекујемо да ће то утицати на резултате. Одговори са оценама 3, 4 и 5 имају прилично једнаку расподелу, са 17, 21 и 22 одговора. Као што видимо на основу матрице конфузије, модел је највише проблема имао са предвиђањем одговора са оценом 1. У тест бази се од укупно седам одговора оцењених са 5 нашло чак шест одговора, а модел је тачно сврстао три (табела 11б).

Добијена прецизност модела је 50%, одзив је 47%, а $F_1 = 0,46$, што је умногоме лошије него код бинарне класификације. Претпостављамо да је оваквим резултатима допринео мали тест скуп података, неједнака расподела ознака у тест и тренинг скупу, као и слична расподела коначног скора одговора у различитим категоријама тачности.

## 7. Закључне напомене

У оквиру овог рада бавили смо се применом латентне семантичке анализе у сврхе процене тачности кратких одговора. У складу са постављеним педагошким циљевима рада, закључили смо да употреба дигиталних лексичких картица у сврхе усвајања вокабулара другог језика даје повољне резултате, нарочито на нижим нивоима знања језика. Како студенти Рударско-геолошког факултета долазе из различитих образовних позадина, те углавном са ниским нивоом почетног знања језика, сматрамо да би увођење система дигиталних лексичких картица које прате наставне материјале у праксу у великој мери помогло студентима да брже напредују, те достигну ниво познавања вокабулара на којима би могли да прате наставу CLIL методом.

Када су у питању методолошки циљеви рада, израда описаног модела помогла нам је да уочимо предности и недостатке нашег приступа. Једна од главних уочених предности је добро моделовање тема на дужем тексту, као и на дефиницијама и одговорима испитаника из домена геологије. Сматрамо да је главни недостатак модела немогућност детектовања тема у веома кратким одговорима, што би се могло превазићи радом на већем скупу података.

Планирани развој овог пројекта подразумева ширење скупа података, затим увођење система за процену тачности писања и процену граматичке исправности одговора. Како бисмо поспешили резултате добијене LSA, покушаћемо да видимо како модел ради са мање димензија и уколико се креирају засебни семантички простори за геолошки и за општи вокабулар. Приликом поређења одговора и лекције, сматрамо да би уместо поређења са целом лекцијом, било боље поредити одговор и фрагмент текста где је одређени геолошки појам описан, тј. где се општи појам који је испитаник требало да дефинише употребљава. Такође, уместо рачунања сличности свих одговора, рачунали бисмо сличности између одговора на исто питање.

Израдом представљеног модела поставили смо темеље за даљи развој система за процену тачности одговора у оквиру дигиталних лексичких картица. Поређењем циљева CLIL метода наставе и исхода употребе дигиталних лексичких картица у настави, закључили смо да је ово технологија која би употпунила уџбеник у припреми проф. др Лидије Беко, а нашу тврдњу додатно је подржао позитиван став према употреби дигиталних лексичких картица у настави језика који су студенти Рударско-геолошког факултета Универзитета у Београду који су слушали предмете Енглески језик 1–4 исказали у претходним истраживањима. Даљим истраживањем настојаћемо да испунимо коначни циљ пројекта —- израду система дигиталних лексичких картица, које ће бити примењиване у настави.

## Захвалност

Како је овај рад настао из мастер рада израђеног у оквиру интердисциплинарног мастер програма Рачунарство у друштвеним наукама при Универзитету у Београду, желела бих да се захвалим проф. др Ранки Станковић за менторство при изради мастер рада, за стрпљење и пружене смернице, знање и подршку. Такође бих желела да се захвалим проф. др Лидији Беко што ме је позвала да свој мастер рад радим на њеном уџбенику у припреми, који ће се користити у настави предмета Енглески 1–4 на Рударско-геолошком факултету, Универзитета у Београду. На крају, захвалила бих се чланици комисије за одбрану мастер рада проф. др Јелени Јовановић, јер су њена питања приликом одбране рада допринела исправкама у овој верзији рада.

## Литература

Ashcroft, Robert John, Robert Cvitkovic, and Max Praver. 2018. "Digital flashcard L2 Vocabulary learning out-performs traditional flashcards at lower proficiency levels: A mixed-methods study of 139 Japanese university students." *The EuroCALL Review* 26 (1): 14–28. https://doi.org/10.4995/eurocall.2018.7881.

Averianova, Irina. 2015. "Vocabulary acquisition in L2: does CALL really help." In *Critical CALL–Proceedings of the 2015 EUROCALL Conference, Padova, Italy,* edited by F. Helm, L. Bradley, M. Guarda, and S. Thouësny, 30–35. https://doi.org/10.14705/rpnet.2015.000306.

Baten, Kristof, Silke Van Hiel, and Ludovic De Cuypere. 2020. "Vocabulary Development in a CLIL Context: A Comparison between French and English L2." *Studies in second language learning and teaching* 10 (2): 307–336.

Beko, Lidija. 2013. "Integrisano učenje sadržaja i jezika (CLIL) na geološkim studijama." Докторска дисертација, Универзитет у Београду, Филолошки факултет.

Beko, Lidija, Ivan Obradović, and Ranka Stanković. 2015. "Developing Students' Mining and Geology Vocabulary Through Flashcards and L1 in the CLIL Classroom." In *The Second International Conference on Teaching English for Specific Purposes Developing students' mining and geology vocabulary through flashcards and L1 in the CLIL classroom.* Faculty of Electronic Engineering, University of Niš.

Chen, Shufeng. 2018. "K-nearest neighbor algorithm optimization in text categorization." In *IOP conference series: earth and environmental science,* 108:052074. 5. IOP Publishing. https://doi.org/10.1088/1755-1315/108/5/052074.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41:391–407.

Đerić, Miloš. 2019. "Doprinos CLIL-a savremenim tokovima nastave stranog jezika." *Philologia* 17 (17): 23–38. ISSN: 1451-5342. https://doi.org/10.18485/philologia.2019.17.17.3.

Dikli, Semire. 2006. "An Overview of Automated Scoring of Essays." *Journal of Technology, Learning, and Assessment* 5 (1). https://ejournals.bc.edu/index.php/jtla/article/view/1640.

Géron, Aurélien. 2022. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media, Inc.

Graesser, Arthur, Peter Wiemer-Hastings, Katja Wiemer-Hastings, Derek Harter, Natalie Person, and Tutoring Research Group. 2000. "Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor." *Interactive Learning Environments* 8:129–148. https://doi.org/10.1076/1049-4820(200008)8:2;1-B;FT129.

Hung, Hsiu-Ting. 2015. "Intentional Vocabulary Learning Using Digital Flashcards." *English Language Teaching* 8:107–112. https://doi.org/10.5539/elt.v8n10p107.

Jhean-Larose, Sandra, Vincent Leclercq, Javier Diaz, Guy Denhiere, and Bernadette Bouchon-Meunier. 2010. "Knowledge evaluation based on LSA : MCQs and free answers." *Stud. Inform. Univ.* 8 (January): 57–84.

Jurafsky, Daniel, and James Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3rd ed. Draft of January 7, 2023. https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Lafourcade, Mathieu, and Virginie Zampa. 2009. "PtiClic: a game for vocabulary assessment combining JeuxDeMots and LSA." In *Advances in Computational Linguistics,* vol. 41 of *Research in Computer Science,* edited by Alexander Gelbuch, 289–298. Center for Computing Research of IPN.

Landauer, Thomas, Darreil Laham, and Peter Foltz. 2003. "Automated scoring and annotation of essays with the Intelligent Essay Assessor." In *Automated essay scoring: A cross-disciplinary perspective,* edited by Mark D. Shermis and Jill C. Burstein, 87–112. Routledge.

Landauer, Thomas K, Darrell Laham, Bob Rehder, and Missy E Schreiner. 1997. "How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans." In *Proceedings of the 19th annual meeting of the Cognitive Science Society,* 412–417.

Landauer, Thomas K., Peter W. Foltz, and Laham Darrell. 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25 (2–3): 259–284.

Lemaire, Benoit, and Philippe Dessus. 2003. "A System To Assess The Semantic Content Of Student Essays." *Journal of Educational Computing Research* 24 (October). https://doi.org/10.2190/G649-0R9C-C021-P6X3.

Li, Baoli, Shiwen Yu, and Qin Lu. 2003. "An Improved k-Nearest Neighbor Algorithm for Text Categorization." In *Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, August 2003.* https://arxiv.org/abs/cs/0306099.

Lifchitz, Alain, Sandra Jhean-Larose, and Guy Denhière. 2009. "Effect of tuned parameters on an LSA multiple choice questions answering model." *Behavior research methods* 41:1201–1209. https://doi.org/10.3758/BRM.41.4.1201.

Ma, Qing. 2009. *Second language vocabulary acquisition.* Vol. 79. Peter Lang.

Martin, Dian I., and Michael W. Berry. 2013. "Mathematical foundations behind latent semantic analysis." In *Handbook of latent semantic analysis,* edited by Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, 35–55. Psychology Press.

Nakata, Tatsuya. 2008. "English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning." *ReCALL* 20 (1): 3–20.

Nation, Paul. 2006. "Vocabulary: Second Language." In *Encyclopedia of Language Linguistics,* 448–454. ISBN: 9780080448541. https://doi.org/10.1016/B0-08-044854-2/00635-0.

Peterson, Leif E. 2009. "K-nearest neighbor." *Scholarpedia* 4 (2): 1883. https://doi.org/doi:10.4249/scholarpedia.1883.

Picca, Davide, Dominique Jaccard, and Gérald Eberlé. 2015. "Natural language processing in serious games: a state of the art." *International Journal of Serious Games* 2 (3): 77–97.

Rahutomo, Faisal, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. "Semantic cosine similarity." In *The 7th international student conference on advanced science and technology ICAST.* https://www.researchgate.net/profile/Faisal-Rahutomo/publication/262525676_Semantic_Cosine_Similarity/links/0a85e537ee3b675c1e000000/Semantic-Cosine-Similarity.pdf.

Spiri, John. 2008. "Online study of frequency list vocabulary with the WordChamp website." *Reflections on English Language Teaching* 7 (1): 21–36.

Urdan, Timothy C. 2005. *Statistics in plain English.* 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Yüksel, H. Gülru, H. Güldem Mercanoğlu, and M. Betül Yılmaz. 2022. "Digital flashcards vs. wordlists for learning technical vocabulary." *Computer Assisted Language Learning* 35 (8): 2001–2017.

# Четврти хакатон великих података о лингвистичким повезаним отвореним подацима

УДК 81'322.2:004.822

Тијана Радовић
tijana.n.radovic@gmail.com
*Универзитет у Београду*
*Београд, Србија*

Ранка Станковић
ranka.stankovic@rgf.bg.ac.rs
*Универзитет у Београду*
*Рударско-геолошки факултет*
*Београд, Србија*

**САЖЕТАК:** Четврти летњи хакатон великих података посвећен лингвистичким повезаним отвореним подацима (*4th Summer Datathon on Linguistic Linked Open Data, SD-LLOD-22*) у оквиру COST акције *NexusLinguarum* одржан је у Шпанији у мају 2022. Школа је окупљала заинтересоване истраживаче, професоре, студенте који су желели да стекну или прошире своја знања из области науке о лингвистичким повезаним подацима. Током школе представљен је спектар тема из области повезаних података, од онтологија, преко интеграције докумената, анотације и алата за обраду текста на природном језику заснованих на повезаним подацима и RDF-ом, до смерница за генерисање RDF-а и објављивање језичких ресурса као и давања увида у значај коришћења повезаних податак у лексикографији и терминологији. Организатор школе био је Политехнички Универзитет у Мадриду а школа је окупила преко 50 учесника како из академских институција тако и из индустрије.

**КЉУЧНЕ РЕЧИ:** лингвистички повезани отворени подаци, анализа сентимента, повезани подаци, RDF.

**РАД ПРИМЉЕН:** 28. новембар 2022.
**РАД ПРИХВАЋЕН:** 30. март 2023.

## 1. Увод

У оквиру COST акције *NexusLinguarum* одржан је Четврти летњи дан података о лингвистичким повезаним отвореним подацима (*4th*

*Summer Datathon on Linguistic Linked Open Data, SD-LLOD-22*), у Шпанији, месту Серседиљи код Мадрида у периоду од 30. маја до 3. јуна 2022. на Политехничком универзитету у Мадриду (*Universidad Politécnica de Madrid*). COST акцију *NexusLinguarum* (CA18209),[1] финансира Европска унија да би се остварила синергија између лингвиста, информатичара, терминолога, језичких стручњака и других заинтересованих истраживача широм Европе, и како би се истражила и проширила област науке о лингвистичким повезаним подацима (*linguistic linked data*) (Dojchinovski et al. 2021).

## 2. Програм и организација школе

Садржај школе SD-LLOD-22 састојао се од предавања из области повезаних података како би учесници добили теоријску основу и упознали се са предметом бављења и методологијом коришћеном у овој области. Потом, одржане су радионице у оквиру којих су се решавали конкретни проблеми из праксе из различитих области рада са подацима, са главним циљем да се учесницима из индустрије и академске заједнице пруже практична знања у области повезаних података примењених на лингвистику. И на крају, кроз практичан рад на реализацији мини пројеката су учесници радили у групама. Кроз сопствене мини пројекте примењивали су научено укључујући генерисање или коришћење лингвистичких повезаних података, а помоћ им је пружао један од неколико доступних ментора, стручњака за специфичне теме. Коначни циљ је био да се учесницима омогући да трансформишу постојеће лингвистичке податке и објаве их као повезане податке на вебу и да развију апликације које користе лингвистичке повезане податке. [2] Слични догађаји су и раније организовани: 2015. и 2017. у Серседиљи (Шпанија) и 2019. у Дагстулу (Немачка).

Различити језички ресурси попут речника, корпуса и база знања широко и отворено су доступни технологијом повезаних података. Међутим, и даље постоји велики број неповезаних ресурса који су похрањени у хетерогеним форматима и са различитим шемама представљања. Додатно, ти ресурси немају стандардне начине програмског приступа подацима (*Application Programming Interface*,

1. https://nexuslinguarum.eu/

2. Информације о догађају и линкови на ресурсе се могу наћи на званичном сајту https://datathon2022.linkeddata.es/

API). Представљање језичких ресурса као повезаних података има бројне предности као што су агрегација и интеграција језичких ресурса коришћењем заједничког модела података (*Resource description framework*, RDF), експлицитно повезивање, публиковање и претраживање на стандардизован начин (SPARQL), побољшано откривање скупова података и услуга, као и коришћење заједничких речника за представљање језичког садржаја и метаподатака. Укратко, повезани подаци омогућавају лакшу претрагу језичких ресурса, њихову доступност, интерпретабилност и поновну употребљивост (Gracia et al. 2022a).

Током школе, учесници су: 1) генерисали и објавили сопствене лингвистичке повезане податке из већ постојећих извора података, 2) примењивали принципе повезаних података и технологије семантичког веба у области језичких ресурса, 3) користили неке од најважнијих модела који се користе за представљање лингвистичких повезаних података, посебно *OntoLex-Lemon* (McCrae et al. 2017), 4) упознали се са токовима обраде текста на природном језику и апликацијама заснованим на повезаним подацима, 5) добили увид у потенцијалне предности повезаних података и могућности њихове примене за специфичне случајеве употребе.

Предавања у оквиру школе обрадила су следеће теме: онтологије и повезани подаци са акцентом на OntoLex-Lemon модел као модел лексикона за онтологије који подржава RDF (*Resource Description Framework*) и OWL (*Web Ontology Language*), две од три фундаменталне технологије коришћене у оквиру семантичког веба (Gracia et al. 2022b). Затим представљена је интеграција докумената, анотације и алати за обраду текста на природном језику засновани на повезаним подацима и RDF-у користећи Web Annotation и NIF (*NLP Interchange Format*) стандарде. Потом су представљене и смернице за генерисање RDF-а и објављивање језичких ресурса као и значај коришћења повезаних података у лексикографији и терминологији. На крају дат је осврт на целокупни значај употребе и примене лингвистички повезаних података.

## 3. *Sentimientos* – један од мини пројеката реализованих у току SD-LLOD-22

Као што смо претходно поменули, један аспект школе подразумевао је практични рад у групама кроз који су учесници кроз сопствене мини пројекте примењивали научено укључујући генерисање или коришћење

лингвистички повезаних података. Овај самостални практични рад замишљен је тако да учесници сходно својим интересовањима предлажу своје „мини пројекте“ који су укључивали конверзију у повезане податке постојећих скупова података, при чему су нарочито подстицани мини пројекти за језике са недовољно развијеним ресурсима. Представници Србије на овом догађају били су студенти докторских студија „Интелигентни системи“ при Универзитету у Београду: Тијана Радовић и Милош Кошпрдић. Заједно са колегом Мартином Алехандром Чајом са Универзитета у Мадриду и ментором Сином Ахмадијем са Универзитета у Болоњи радили су на пројекту *SD-LLOD-22 Sentimientos* чији је циљ био да обогати, конвертује и публикује лексикографске податке са информацијама о поларитету речи у формату повезаних података. За српски језик коришћен је лексикон *Senti-Pol-sr* (Stanković et al. 2022), а за немачки језик лексикон *PolArt* (Klenner, Fahrni, and Petrakis 2009). Структура оба лексикона је сасвим једноставна: лема и придружен поларитет исказан кроз колоне позитивно и негативно са дискретним вредностима 0 и 1.

Пројекат се састојао из три дела:

1) почевши од листе лема, прикупљени су лексикографски подаци из базе *Dbnary* (Sérasset 2012) користећи SPARQL упите. Креиран је шаблон онтологије: шема лексичких записа која се састојала од лексичког записа (*LexicalEntry*), врсте речи (*partOfSpeech*), леме (*lemma_label*), лексичког значења (*lexicalSense*) и поларитета (*hasPolarity*), што се може видети у коду који следи (плавом бојом су обележени делови који се мењају). Коришћењем представљеног обрасца је даље сваки запис из улазних речника трансформисан у RDF облик.

```
rdf_template = """
<:le_uri> a ontolex:lexicalEntry;
   lexinfo:partOfSpeech <:pos_uri>;
   rdfs:label 'lemma_label'@language_tag;
   ontolex:lexicalSense :senses_uris.

<:lemma_opinion> marl:describesObject <:le_uri>;
   marl:hasPolarity polarity.
"""
```

2) Да би се произвела такозвана *ttl* датотека, подаци су допуњени коришћењем SPARQL упита над лексиконом *Dbnary* како би се преузео лексички запис, врста речи и значење за конкретне речи из улазних

речника. У наставку приказујемо SPARQL упит за допуну постојећих података:

```
query = (
  "\n"
  " PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>\n"
  " PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>\n"
  " PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>\n"
  " \n"
  " SELECT ?le ?pos ?sense \n"
  " WHERE {\n"
  " ?le a ontolex:LexicalEntry .\n"
  " ?le rdfs:label \"xxxxx\"@<language2>.\n"
  " ?le lexinfo:partOfSpeech ?pos.\n"
  " ?le ontolex:sense ?sense\n"
  " }"
    "")
```

3) Упит се прослеђује приступној тачки *Dbnary*, у овом случају на следећи начин:

```
sparql = SPARQLWrapper( "http://kaiko.getalp.org/sparql")
sparql.setReturnFormat(JSON)

def run_query(word, language, language_):
    word_query = query.replace("xxxxx", word)
    word_query = word_query.replace("<language2>", language)
    sparql.setQuery(word_query)
    return sparql.queryAndConvert()["results"]["bindings"]
```

Пример лексичког записа који садржи лему, врсту речи и поларитет осећања би у овој трансформацији изгледао како је приказано на слици 1.

У овом мини пројекту коришћени су поменути српски и немачки лексикони сентимента и *Dbnary* али се приступ може проширити и на друге врсте лексикона.

У циљу интегрисања произведених алата у корисничку апликацију, произведени подаци су похрањени у базу *GraphDB*[3] и над њима је направљена мала апликација коришћењем програма Јава скрипт за претраживање похрањених података.

3. GraphDB

```
### word: hrana
<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>
  a ontolex:lexicalEntry;
  lexinfo:partOfSpeech <http://www.lexinfo.net/ontology/2.0/lexinfo#noun>;
  rdfs:label 'hrana'@sr;
  ontolex:lexicalSense
  <http://kaiko.getalp.org/dbnary/ell/__ws_1__srp__hrana__ουσιαστικό__1>.

<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1/opinion>
  marl:describesObject
  <http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>;
  marl:hasPolarity marl:positive.
```

**Слика 1.** Пример лексичког записа.

Коришћени и произведени подаци, као и развијени код су јавно доступни,[4] а додатна верзија је допуњена врстама речи из *Српског морфолошког речника* (Krstev 2008) и коришћењем лексичке базе *Лексимирка* (Stanković et al. 2018). Датотека *Senti-Pol-sr.ttl* доступна је за директан приступ и преузимање на страници Друштва за језичке ресурсе и технологије JePTex,[5] као и за претраживање коришћењем SPARQL језика на приступној тачки JePTex-а која је имплементирана коришћењем алата Fuseki.[6]

## 4. Закључак

Четврти летњи хакатон великих података о лингвистичким повезаним отвореним подацима нама као учесницима је био драгоцен начин за повезивање и размену знања, идеја и искустава са истраживачима сличних интересовања како из академских институција тако и из индустрије. Садржај је био разноврстан, са темама које су прилагођене знању полазника у различитим секцијама, тако да смо успели да проширимо раније стечена знања и добијемо смернице за будућа истраживања. Предавања теоријског типа помогла су нам да боље разумемо методологију и концепте повезивања, као и шеме

4. SD-LLOD-22 Sentimientos
5. http://llod.jerteh.rs
6. http://fuseki.jerteh.rs//dataset/Senti-Pol-sr/query

и стандарде репрезентације података. На радионицама, радом на конкретним проблемима упознали смо се са различитим технологијама у области повезаних података. Стицање знања и вештина током пет дана трајања школе је омогућило да генеришемо и објавимо сопствене лингвистичке повезане податке из већ постојећих извора података, што ће нам омогућити даља истраживања у овом правцу: генерисање сличних скупова података, надградњу кроз обогаћивање и проширивање публикованог лексикона. Такође, упознали смо се са токовима обраде текста на природном језику и апликацијама заснованим на повезаним подацима и уверили се у могућности њихове примене. Рад на мини пројекту *Sentimientos* допринео је обогаћењу ресурса на српском језику конвертовањем и публиковањем листе речи са придруженим информацијама о врсти речи и поларитетом у формату повезаних података (у виду *ttl* датотеке и на приступној тачки), а тај ресурс се може користити за будућа истраживања везана за одређивање сеинтимента текста, али и за објављивање лингвистичких података на сличан начин. Повезивање лексикона са примерима употребе који би били публиковани у *NIF* формату један је од првих задатака којим ћемо се позабавити у наредном периоду.

## Захвалност

Представљени рад је подржала COST акција *CA18209 – NexusLinguarum "European network for Web-centred linguistic data science"*. Аутори се захваљују Милошу Кошпрдићу на учешћу у припреми скупа података.

## Литература

Dojchinovski, Milan, Julia Bosque Gil, Jorge Gracia, and Ranka Stanković. 2021. "EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School." *Infotheca* 21 (1): 113–120. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.1.7.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022a. *SD-LLOD-22 Course Slides,* October. https://doi.org/10.5281/zenodo.7197674.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022b. *SD-LLOD-22 Materials,* October. https://doi.org/10.5281/zenodo.7197696.

Klenner, Manfred, Angela Fahrni, and Stefanos Petrakis. 2009. "Polart: A robust tool for sentiment analysis." In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009),* 235–238.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology, University of Belgrade.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Sérasset, Gilles. 2012. "Dbnary: Wiktionary as a LMF based Multilingual RDF network." In *Proc. of the 8th Int. Conf. LREC'12,* edited by Nicoletta Calzolari et al. ELRA. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data,* 31–38.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries – from File System to *lemon* Based Lexical Database." In *Proc. of the 11th Int. Conf. LREC 2018 - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), Miyazaki, Japan, May 7-12, 2018,* edited by John P. et al McCrae. ELRA. http://lrec-conf.org/workshops/lrec2018/W23/pdf/3_W23.pdf.

# Теоријски основ и креирање паралелног корпуса: оцртавање нових перспектива у обради библиотечког материјала

Драгана Столић
stolic@unilib.rs
*Универзитетска библиотека „Светозар Марковић” Београд, Србија*

**РАД ПРИМЉЕН:** 24. новембар 2022.
**РАД ПРИХВАЋЕН:** 31. март 2023.

## 1. Увод

Публикација Јелене Андоновски *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семнатички веб*[1] придружује се вредној издавачкој колекцији Универзитетске библиотеке „Светозар Марковић“, покренутој 2016. године под називом „Хомилије“, намењеној публиковању научних монографија, које махом чине прерађене одбрањене докторске дисертације аутора из друштвено-хуманистичких наука, односно дигиталне хуманистике. И овде је реч о прилагођеном тексту дисертације *Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литератног корпуса*, коју је ауторка одбранила на Филолошком факултету у Београду 2020. године.

Публикација представља високопрецизну систематизацију знања и умећа везаних за обраду, опис и истраживање језичких корпуса како би се створиле претпоставке за потоње лексичке, терминолошке и семантичке анализе. Поступно и логично, дело нас води ка разумевању овог специфичног сегмента библиотекарства и информатике, оцртава модел на који начин треба приступити таквом задатку, не заборављајући да укаже на путање даљих истраживања. Уопштено говорећи, овакве публикације помажу запосленима у библиотечко-информационој делатности да јасније сагледају читав домен истраживања који је,

1. Јелена Андоновски: *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семантички веб*, Београд: Универзитетска библиотека „Светозар Марковић“, 2021.

условљен технолошким напретком и развојем нових алата, омогућио аналитичко задирање у публикације саме, придружујући их доступним надређеним целинама и обезбеђујући раније немогуће приступе књижевним делима и писцима. Као што класично библиотекарство подразумева значајан број разних активности, прикупљања, обраде и презентације публикација како би се кориснику омогућило да дође до жељеног извора информација, тако и нове информационе технологије допуштају анализу и обраду језичке грађе како би се разумела природа језика и његовог коришћења и како би се утабали путеви за лакше превођење и разумевање других језика.

## 2. Пут ка истраживању корпуса

Иако публикација има више поглавља, условно се могу уочити две крупније целине: једна, коју можемо назвати теоријским односно *појмовним полазиштем* и другу, која је укључила примену свега претходно наведеног на одређени корпус (СрпНемКор), што је подразумевало употребу одређених алата, спровођење претраживања и анализу добијених резултата.

Поменути први део најпре одређује појам *корпуса*, као основе *корпусне лингвистике*, а у оквиру тога – *паралелних корпуса*. У делу се пружа јасна типологија корпуса, који могу бити једнојезични и вишејезични, који се даље могу поделити на *паралелне* и *упоредне*. Паралелни су састављени од текстова на изворном језику и њихових превода на један или више циљаних језика и као такви значајни су за учење страних језика, различита превођења, нарочито машинска, затим терминолошка и лингвистичка истраживања и сл. Истакнуто је да је основни елемент двојезичног корпуса *битекст* или *паралелизовани* текст, где се повезивање текста и његовог превода спроводи на различитим нивоима – од читавог документа, преко поглавља, пасуса до реченице и речи, са циљем *упаривања текстова (паралелизације)* или стварања веза између варијанти јединица превођења и формирања скупа јединица превођења („Translation Unit“).

После осврта на прве видљиве покушаје формирања паралелних корпуса, у раду је читаво поглавље посвећено постигнућима у Србији на овом пољу, пре свега раду Друштва за језичке ресурсе и технологије - JePTex[2]. На тај начин је пружен кратак и сажет, али веома информативан преглед резултата вишегодишњих пројеката.

2. JePTex

Поступак израде паралелних корпуса представљен је у свим корацима, чиме се значајно олакшава разумевање опсега оваквог приступа, где посебну тежину имају анотације, највише оне које укључују коришћење проширивог језика за обележивање (XML – eXtensible Markup Language) и све више присутну, али чини се недовољно прихваћену, иницијативу за кодирање текста – TEI (Text Encoding Initiative). Изабраним сегментима се придодају обележја (етикете) која одређеним речима додељују додатне дескрипторе за врсту речи, канонски облик или морфолошку категорију. Припреми материјала придружује се потом приказ предуслова за проналажење информација где кључну улогу имају правилно додељени метаподаци. Појам „метаподатак“, који практично узима превагу у савременом библиотекарству у односу на класичне библиографске податке, образложен је пажљиво у свим облицима њиховог испољавања (описни, структурални, административни) и приказом основних схема (Dublin Core, METS, MODS).

На самом крају овог, како је речено, фундаменталног сегмента, у којем су формулисане дефиниције и оцртан читав систем средстава неопходних када је у питању стварање једног корпуса, као врхунац и најасложенији појам образлаже се *семантички веб*. Веома стручно, али довољно разумљиво, онај део рада се хвата укоштац са иницијативом „отворених повезаних података“ (Linked Open Data) настојећи да појасни функционисање веба као глобалне базе података чија се структура више не заснива на статичним страницама, већ, кроз механизам за увезивање података на нивоу њиховог значења, тј. на садржајима повезаним према семантичким карактеристикама.

## 3. Израда српско-немачког паралелног корпуса

После исцрпног приказа и анализе алата и ресурса који се могу користити у стварању паралелног копруса, други део студије (поглавље 6) бави се апликацијом свега наведеног разматрајући могућности, домете или ограничења неких ресурса на датом материјалу. За формирање корпуса коришћено је седам романа различитих српских писаца и исти број дела различитих писаца са немачког говорног подручја, руководећи се критеријумима као што су: награђиваност аутора, доступност дела на оба језика у библиотекама Србије, популарност дела и његова обимност.

Приказани су кораци у стварању корпуса од дигитализације материјала, примене оптичког препознавања текста, структурне

анотације и, најзад, до постављања корпуса у оквир дигиталне библиотеке *Библиша*[3] и обезбеђивања услова за квалитетно претраживање информација, као и за укључивање обрађеног корпуса у мрежу отворених лингвистичких података. Читав приказ јесте аутентична илустрација свих нужних радњи, али и изазова, проблема и ограничења чиме резултати овог рада не губе на значају.

## 4. Улога библиотека

Ова студија је практично, кроз интеграцију више елемената, и сама креирала један нови „корпус“ појмова и значења који на другачији начин сагледавају текст. У том смислу, чини се да класична библиотечко-информациона делатност нема нужно додирних тачака са таквим приступом. Међутим, посебан квалитет овог рада управо јесте доследно истицање незаобилазне улоге библиотека, које, и поред доминантно „несемантичког“ приступа, морају да чине део једног оваквог процеса, не само због дигитализације где најактивније учествују, или чињенице да поседују грађу која је полазиште у креирању корпуса. Наиме, управо библиотеке, у којима се већ креирају различити метаподаци и исписи, поседују предуслове да учине одговарајући искорак према значењском повезивању добијених података и обезбеђивању потпунијих информација за кориснике.

3. Библиша