**Инфо***тека* Часопис за дигиталну хуманистику
Београд, год. 24, бр. 1(2024), фебруар 2025.

# 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 УРЕДНИЦИ:
**др Александра Тртовац** aleksandra@unilib.rs
**проф. др Ранка Станковић** ranka.stankovic@gmail.com

УРЕДНИЦИ ЕЛЕКТРОНСКОГ ИЗДАЊА:
**др Јелена Гавриловић** andonovski@unilib.rs
**Ивана Гавриловић** gavrilovic@unilib.rs

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

ИЗРАДА ВЕБ ПОРТАЛА ЧАСОПИСА:
**др Јелена Гавриловић**
**Ивана 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:
**Доц. 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*

Часопис излази два пута годишње

**Инфо***тека* **Год. 24, бр. 1(2024)**

# Садржај

# Нови језички модели за српски језик

УДК 811.163.41'322.2

**САЖЕТАК:** У раду ће укратко бити приказан историјат развоја језичких модела за српски језик који су засновани на трансформерској архитектури. Биће представљено и неколико нових модела за генерисање и векторизацију текста, обучених на ресурсима Друштва за језичке ресурсе и технологије. Десет одабраних модела за векторизацију српског језика, међу којима су и два нова модела, биће упоређена на четири задатка обраде природног језика. Анализираћемо који су модели најбољи за изабране задатке, како величина модела и величина скупа за обучавање утичу на њихове перформансе на тим задацима и шта је потребно за обучавање најбољих модела за српски језик.

**КЉУЧНЕ РЕЧИ:** језички модели, српски језик, векторизација, обрада природног језика.

**РАД ПРИМЉЕН:** 27. јануар 2024.
**РАД ПРИХВАЋЕН:** 21. фебруар 2024.

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
ORCID: 0000-0003-4811-8692
*Универзитет у Београду*
*Рударско-геолошки факултет*
*Београд, Србија*

## 1. Увод

Почетком двадесет и првог века, дошло је најпре до наглог пораста количине доступних текстуалних података, а потом и до наглог раста рачунарске моћи, што је покренуло талас истраживања заснованих на идеји дубоког учења (*deep learning*) (LeCun, Bengio, and Hinton 2015). У случају обраде природних језика, истраживања кулминирају појавом архитектуре трансформера (Vaswani et al. 2017), заснованој на употреби енкодера, чија је главна намена анализа текста, и декодера, који су задужени за синтезу текста. Први изразито популаран модел овог типа био је *BERT*[1] (Devlin et al. 2018), заснован искључиво

1. *Bidirectional Encoder Representations from Transformers* – двосмерно кодирање репрезентације из трансформера

на трансформерском енкодеру. Овај модел направио је велики помак у обради природних језика, пре свега на задацима заснованим на векторизацији текста. Његове варијације, *RoBERTa*[2] (Liu et al. 2019) и *DeBERTa*[3] (He et al. 2020) и данас постижу најбоље резултате на задацима угнежђивања речи (*word embedding*), анотације речи (нпр. обележавање врсте речи и препознавање именованих ентитета) и класификације реченица и докумената. Са друге стране, појављивање модела *GPT* (генеративни предобучени трансформер) (Radford et al. 2018) и *GPT-2* (Radford et al. 2019) популаризовало је језичке моделе засноване на траснформерском декодеру, а ова група модела се данас најбрже развија. Модели који комбинују употребу енкодера и декодера, као што су, на пример *BART* (Lewis et al. 2020) и *T5* (Raffel et al. 2020), остају недовољно запажени упркос изванредним резултатима које постижу на задацима трансформације текста, као што су машинско превођење, сумаризација и прилагођавање стила.

### 1.1 Преглед објављених модела за српски језик

Језички модели засновани на архитектури трансформера направили су продор у српски језик путем вишејезичних модела, најпре кроз *MBERT*[4] (Devlin et al. 2018), а потом и кроз *XLM-RoBERTa* модел[5] (Conneau et al. 2019), за чије је обучавање коришћено око четири милијарде токена из текстова писаних на српском или другом сродном језику (хрватски, босански). Потоњи модел објављен је у децембру 2019. године у две варијанте – *base* (279 милиона параметара) и *large* (561 милион параметара). И данас се, као један од највећих модела за векторизацију, употребљава у обради српског језика и притом остварује добре резултате, поготово након дообучавања.

Почетком 2021. године, на платформи *Huggingface*[6] објављен је модел под називом *БЕРТић* (*classla/bcms-bertic*) (Ljubešić and Lauc 2021), базиран на архитектури *ELECTRA* (Clark et al. 2020), са 110 милиона параметара, обучаван на корпусу од преко осам милијарди токена, босанског (800 милиона), хрватског (5,5 милијарди), црногорског (80 милиона) и српског језика (2 милијарде).

2. *Robustly Optimized BERT* – Робусно оптимизовани *BERT*
3. *Decoding-Enhanced BERT* – *BERT* проширен декодирањем
4. *Multilingual BERT* – вишејезични *BERT*
5. *Cross-lingual Language Model* – Међујезички језички модел
6. Huggingface, највеће веб чвориште за објављивање језичких модела.

Касније исте године направљени су и објављени први специфични модели за српски језик у оквиру једног ширег језичког истраживања за македонски језик (Dobreva et al. 2022). Прецизније, објављена је српска верзија *RoBERTa-base* модела, *macedonizer/sr-roberta-base* (120 милиона параметара) и српска верзија *GPT2-small* модела, *macedonizer/sr-gpt2* (130 милиона параметара). Оба модела су обучавана на корпусу српске Википедије и подржавају само ћирилично писмо.

Недуго потом предузет је сличан подухват, при којем је обучено пет *RoBERTa-base* модела за српски језик (Cvejić 2022). Првобитни модел *Andrija/SRoBERTa* имао је 120 милиона параметара и обучаван је на малом корпусу од 18 милиона токена познатом под именом *Leipzig* (Biemann et al. 2007), док су потоња четири модела имала по 80 милиона параметара, при чему је сваки обучаван на све већем корпусу. За модел *Andrija/SRoBERTa-base* додат је корпус *OSCAR* (Suárez, Sagot, and Romary 2019) (220 милиона токена), за модел *Andrija/SRoBERTa-L* је поред њега додат и *srWAc* (Ljubešić and Klubička 2014) (490 милиона токена), за модел *Andrija/SRoBERTa-XL* је уз претходне додат и део корпуса *cc100-hr* (21 милијарда токена) и *cc100-sr* (5,5 милијарди токена) (Wenzek et al. 2020), док су за модел *Andrija/SRoBERTa-F* сви поменути корпуси коришћени у целости.

Крајем 2022. године објављена су три експериментална генеративна модела за српски језик (Škorić 2023). Контролни модел *procesaur/gpt2-srlat* био је поново заснован на *GPT2-small* архитектури, имао је 138 милиона параметара и обучен је на исечку корпуса Друштва за језичке ресурсе и технологије (260 милиона токена) (Krstev and Stanković 2023). Друга два модела – *procesaur/gpt2-srlat-sem* и *procesaur/gpt2-srlat-synt*, настала су дообучавањем контролног модела коришћењем два специјално припремљена корпуса са циљем засебног моделовања семантике, односно, синтаксе текста. Три модела су потом употребљена за експеримент комбиновања језичких модела на задатку класификације реченица (Škorić, Utvić, and Stanković 2023).

Почетком наредне године, истраживачи са Универзитета у Нишу објавили су модел *JelenaTosic/SRBerta* (75 милиона параметара) који је такође заснован на *RoBERTa-base* аархитектури, а обучаван је помоћу корпуса *OSCAR* (Suárez, Sagot, and Romary 2019).Занимљиво је да је овај модел, као и његова друга верзија (*nemanjaPetrovic/SRBerta*, 120 милиона параметара), пре објављивања дообучен над текстовима из домена права (Bogdanović, Kocić, and Stoimenov 2024). У периоду између објављивања ових модела, објављен је и *aleksahet/xlm-r-squad-sr-*

*lat* (Cvetanović and Tadić 2023), први модел за одговарање на питања на српском језику, настао прилагођавањем *XLM-RoBERTa* модела помоћу скупа података *SQuAD* (Rajpurkar, Jia, and Liang 2018), преведеног на српски језик.

Средином 2023. године објављена су још два генеративна модела заснована на ГПТ архитектури. Оба модела су обучавана над истим скупом података: над корпусима Друштва за језичке ресурсе и технологије (Krstev and Stanković 2023), докторским дисертацијама преузетим са платформе НАРДУС,[7] корпусу јавног дискурса српског језика Института за српски језик САНУ под називом *PDRS* (Wasserscheidt 2023) и додатним јавно доступним корпусима са веба, као што су поменути *srWAc* (Ljubešić and Klubička 2014) и *cc100-sr* (Wenzek et al. 2020). Укупан број токена у овом скупу података броји око четири милијарде. Већи модел – *jerteh/gpt2-orao*[8] броји 800 милиона параметара, заснован је на архитектури *GPT2-large* и представља тренутно највећи доступни модел предобучен за српски језик. Мањи модел – *jerteh/gpt2-vrabac*[9] броји 136 милиона параметара и заснован је на архитектури *GPT2-small*. Оба модела су обучавана коришћењем рачунарских ресурса Националне платформе за вештачку интелигенцију Србије. Осим корпуса за обучавање, ова два модела деле и речник токена и токенизатор, специјално опремљен да упарује ћирилична и латинична слова, омогућујући им равноправну подршку.

Након објављивања генеративног модела од 800 милиона (*jerteh/gpt2-orao*) параметара, фокус се полако помера на дообучавање великих модела коришћењем текстова на српском језику. Тако су објављена два модела заснована на архитектури *Alpaca* (Taori et al. 2023), *datatab/alpaca-serbian-3b-base* (3 милијарде параметара) и *datatab/alpaca-serbian-7b-base* (7 милијарди параметара), а најављено је и објављивање још једног модела исте величине, заснованог на архитектури *Mistral-7b* (Jiang et al. 2023), који је обучаван на хрватским, босанским и српским текстовима (11,5 милијарди токена). Исти корпус коришћен је и за дообучавање *XLM-RoBERTa-large* модела у циљу поређења перформанси дообучаваних модела у односу на моделе обучаване од почетка (*од нуле*). Нови модел објављен је под именом *classla/xlm-r-*

7. НАРДУС – Национални репозиторијум докторских дисертација са свих универзитета у Србији.
8. jerteh/gpt2-orao
9. jerteh/gpt2-vrabac

*bertic* (Ljubešić et al. 2024) и броји 561 милион параметара, колико има и оригинални *XLM* модел.

Коначно, на скупу података над којим су обучавани *jerteh/gpt2-orao* и *jerteh/gpt2-vrabac*, обучена су још два модела за векторизацију текста. Већи модел, *jerteh/Jerteh-355*,[10] заснован је на *RoBERTa-large* архитектури и броји 355 милиона параметара, док је мањи модел, *jerteh/Jerteh-81*,[11] заснован на *RoBERTa-base* архитектури и броји 81 милион параметара. Као и код модела *jerteh/gpt2-orao*, циљ је да се модели обуче на што квалитетнијем корпусу текстова. У овом раду биће представљено испитивање перформанси ових модела и њихово поређење са перформансама других одабраних модела како би се установило њихово место у хијерархији језичких модела за векторизацију текста на српском језику.

### 1.2 Поставка експеримента

У претходном одељку указано је на постојање већег броја вишејезичних модела који, у мањој или већој мери, подржавају обраду српског језика, као и на двадесетак модела који су припремљени специјално за обраду српског језика. Објављени модели се међусобно разликују према неколико особина: породици (архитектури) модела и броју његових параметара, речнику, односно токенизатору, на којем се заснивају, скупу коришћеном за њихово обучавање, задатку на којем је модел обучаван и дужини обучавања. Треба напоменути да неке од ових информација недостају за неке моделе, али и да су неке доступне информације (пре свега особине скупа коришћеног за обучавање) непроверљиве.

У наставку, рад ће се фокусирати на десет одабраних модела за векторизацију (општег типа). Основне информације о тим моделима биће представљене у одељку 2., експеримент поређења њихових перформанси на четири припремљена задатка приказаћемо у одељку 3., а резултати експеримената биће приказани и размотрени у одељку 4.. Напослетку, у одељку 5., биће предложен процес обучавања нових модела за српски језик. Овај рад неће разматрати генеративне моделе услед недостатка поузданог (али аутоматског) механизма за мерење њихових перформанси. Још није објављен ниједан енкодер-декодер модел специјално развијен за српски језик.

10. jerteh/Jerteh-355
11. jerteh/Jerteh-81

## 2. Одабрани модели за векторизацију текста

За потребе овог рада, од претходно поменутих модела (одељак 1.) одабрано је десет за детаљнију анализу. У тих десет улазе најпре четири *SRoBERTa* модела, који су врло погодни за овај експеримент јер се разликују искључиво по скупу података за обучавање. Даље, ту су најстарији модел, *classla/bcms-bertic* и најновији модел, *classla/xlm-r-bertic*, које је објавио центар за јужнословенске језике *CLASSLA*, као и два најпопуларнија вишејезична модела *xlm-roberta-base* и *xlm-roberta-large*. Коначно, ту су два модела која се први пут представљају у овом раду, модели *jerteh-81* и *jerteh-355*, који су обучени над ресурсима Друштва за језичке ресурсе и технологије.

Основне карактеристике ових десет модела приказане су у табели 1.

У приложеној табели, као и из описа модела у претходном одељку, види се да је најпопуларнија архитектура *RoBERTa* (6 од 10 одабраних модела), а додатна три модела заснована су на блиској, *XLM-RoBERTa* архитектури. Преостали модел *bcms-bertic*, заснива се на ELECTRA архитектури и једини је од одабраних који није обучаван на задатку моделовања маскираног језика (предвиђања делова текста маскираних иза неке специјалне етикете).

Величина одабраних модела варира од 80 (за четири *SRoBERTa* модела) па до преко 560 милиона параметара (за моделе засноване на *XLM-RoBERTa-large*). Величина скупа за обучавање варира од 500 милиона за модел 1 (*SRoBERTa-base*), па до чак 11,5 милијарди токена за модел 6 (*classla/xlm-r-bertic*), при чему треба напоменути да није обучаван од нуле, већ је у питању *xlm-roberta-large* модел дообучен за хрватски, босански и српски језик. Само четири од десет модела обучавана су искључиво над корпусом српских текстова. У питању су модели 1, 2, 9 и 10, тј. прва два *SRoBERTa* модела, *jerteh/jerteh-81* и *jerteh/jerteh-355*.

Битно је напоменути и да десет приказаних модела користе само четири различита речника токена, односно токенизатора:

$X_1$ *SRoBERTa* токенизатор - прва 4 модела;
$X_2$ *bertic* токенизатор - модел број 5;
$X_3$ *XLM-R* токенизатор - модели 6 до 8;
$X_4$ *jerteh* токенизатор - последња 2 модела (9 и 10).

| редни број | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Идентификатор | Andrija/SRoBERTa-base | Andrija/SRoBERTa-L | Andrija/SRoBERTa-XL | Andrija/SRoBERTa-F | classla/bcms-bertic | classla/xlm-r-bertic | xlm-roberta-base | xlm-roberta-large | jerteh/jerteh-81 | jerteh/jerteh-355 |
| Речник токена | SRoBERTa | | | | bertic | XLM-R | | | jerteh | |
| Архитектура | RoBERTa | | | | ELE. | XLM-R | | | RoBERTa | |
| Величина модела | 80 | | | | 110 | 561 | 279 | 561 | 81 | 355 |
| Величина скупа | 500 | 1.000 | 3.750 | 5.700 | 8.400 | 11.500 | 4.000* | | 4.000 | |
| Српски | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Хрватски | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | |
| Босански | | | | | ✓ | ✓ | ✓ | ✓ | | |
| Црногорски | | | | | ✓ | | ✓ | ✓ | | |

**Табела 1.** Десет одабраних модела за векторизацију текста на српском језику и њихове особине: речник токена на којем су засновани, архитектура модела, величина модела изражена у милионима параметара и величина скупа изражена у милионима токена. Подаци су преузети са платформе *HuggingFace*. *Величина скупа за обучавање код модела 7 и 8 (*xlm-roberta-base*, *xlm-roberta-large*) односи се на део скупа на српском, хрватском или другом сродном језику. Доњи део табеле приказује на којем од ових језика су обучавани модели.

## 3. Поставка евалуације перформанси модела

Десет одабраних модела је евалуирано на четири засебна задатка како би се упоредиле њихове перформансе:

$T_1$ Моделовање маскираног језика (погађање недостајућих токена);
$T_2$ Израчунавање (семантичке) сличности између реченица;
$T_3$ Обележавање врстом речи;
$T_4$ Препознавање именованих ентитета.

Прва два задатка припадају групи такозваних узводних задатака (*upstream*), тј. задатака који користе моделе у основном стању, док друга два задатка припадају групи низводних задатака (*downstream*) јер захтевају да се модели фино подесе (*fine-tune*) и тестирају на специјално припремљеном скупу података.

### 3.1 Евалуација модела на узводним задацима

Као што је већ поменуто, за узводне задатке није неопходно прилагођавање модела, те је потребно само припремити скупове за тестирање.

Како би се спровела евалуација модела на задатку моделовања маскираног језика ($T_1$) најпре је припремљен посебан скуп података у виду текстова у којима су у свакој реченици по један насумично изабрани токен маскирани – по један токен је сакривен, на пример иза маске `<MASK>`. За текстуалну грађу коришћена су четири извора:

$Y_1$ *Дечко*, српски превод романа *Подросток* од Достојевског;
$Y_2$ *Младић*, алтернативни превод *Дечка*;
$Y_3$ *Пут око света за 80 дана*, српски превод романа Жила Верна;
$Y_4$ *Пут око свијета у 80 дана*, хрватски превод романа Жила Верна.

Прва два извора нису коришћена за обучавање ниједног модела, док су друга два већ дуго доступни на вебу (Vitas et al. 2008) и стога су вероватно коришћени за обучавање већине, ако не и свих, наведених модела. Како неки модел не би био у посебној предности, текстови су токенизовани коришћењем сва четири токенизатора ($X_1$ до $X_4$), потом маскирани, а онда је пред сваким од модела био задатак да одмаскира свих шеснаест припремљених текстова (четири извора токенизована и маскирана на четири начина). У свакој реченици био је маскиран по

један токен, а модели су током евалуације за његово место нудили по три кандидата. За процену резултата теста узимана је мера тачности на овом задатку, при чему се као погодак рачунало свако одмаскирање код кога је маскирани, то јест тражени, токен био у скупу кандидата које је модел понудио за задату реченицу.

За потребе евалуације на задатку израчунавања сличности између реченица ($T_2$), коришћене су реченице из истих романа, то јест два пара паралелизованих романа ($Y_1$ и $Y_2$, односно $Y_3$ и $Y_4$), али припремљене као триплети. С обзиром на то да су романи претходно паралелизовани на нивоу реченице, било је лако направити парове реченица које имају исто значење. Сваки триплет је употпуњавала додатна реченица из романа парњака, која дели што је више могуће токена са контролном реченицом и има сличну дужину, али је повучена из неког другог места у тексту. Пример триплета:

1. "Zaista, ko ne bi obišao svet i za manju cenu?" (контролна реченица, $Y_3$: *Пут око света за 80 дана*)
2. "Doista, nije li i za manje od toga vrijedno izvršiti put oko svijeta?" (парњак, $Y_4$: *Пут око свијета у 80 дана*)
3. "He! he! pa konačno zašto ne bi uspio?" (лажни парњак, $Y_4$: *Пут око свијета у 80 дана*)

Задатак модела био је да у задатим триплетима препознају прави парњак (сличност између прве и друге реченице треба да буде већа него између прве и треће), а за процену резултата теста узимана је тачност на том задатку. Да би се израчунао реченични вектор, модел најпре додели векторске вредности сваком токену у реченици, а потом се вредност тих вектора усредњава како би се добила векторска репрезентација реченице. Сличност између двеју реченица израчунава се као разлика броја 1 и косинусне удаљености израчунатих реченичних вектора.

### 3.2 Евалуација модела на низводним задацима

За потребе евалуације модела преостала два предвиђена задатка, модели су дообучени и тестирани на посебним скуповима података. За обележавање врстом речи ($T_3$) коришћен је јавно доступни скуп *SrpKor4Tagging* (Stanković et al. 2020) (триста педесет хиљада обележених токена), док је за препознавање именованих ентитета ($T_4$) коришћен други јавно доступни скуп *SrpELTeC-gold* (Šandrih Todorović et al. 2021). У оба случаја, модели су дообучени на 90% обележених реченица из

сваког скупа и тестирани на преосталих 10%. Како је реч о проблему вишекласне класификације, за процену резултата ова два задатка узете су $F_1$-мере остварене приликом класификације над реченицама из скупа за тестирање.

## 4. Резултати евалуације

Резултати првог теста ($T_1$), тј. просечна тачност одабраних модела на задатку допуњавања недостајућег токена у сваком од шеснаест припремљених текстова, приказани су у табели 2.

У приложеним резултатима је може се уочити супериорност новог модела, *jerteh-355*, који остварује бољи резултат од осталих модела у тринаест од шеснаест случајева, односно бољи или исти резултат ($\pm 1\%$) у петнаест од шеснаест случајева. Осим тога, у девет од дванаест случајева модел *jerteh-355* надмашује друге моделе чак и када обрађује текст маскиран токенизатором тих модела. Једини модел који успева да га надмаши у два случаја је *SRoBERTa-F*, који се најчешће показује као најбољи у обради извора ($Y_4$) писаног на хрватском језику, који је у великом проценту био укључен у његов скуп за обучавање. Ипак, у просеку, његова тачност на овом тесту је нижа и од оне коју остварује други нови модел, *jerteh-81*. Модел 5 (*classla/bcms-bertic*), који, за разлику од осталих, није обучаван на задатку моделовања маскираног језика, није био укључен у евалуацију на овом задатку јер би био у неповољном положају.

Резултати теста израчунавања сличности између реченица ($T_2$) приказани су у табели 3. Вредности приказују тачност модела при препознавању реченица са истим/сличним значењем у триплетима екстрахованим из два српска превода истог романа, $Y_1$ і $Y_2$ (први ред вредности), из српског и хрватског превода истог романа, $Y_3$ и $Y_4$ (други ред), и просечну тачност (трећи ред).

Резултати за први низ триплета су веома добри за неколико модела: *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* и *jerteh-355* остварују сличну тачност од преко 95%. Модел *SRoBERTa-XL* остварује најбоље резултате, уз малу маргину, али и најбољи резултат за други низ триплета који садржи реченице на хрватском језику (92% тачности), па самим тим има и најбољи просечни резултат на овом задатку. Једини други модел који остварује тачност од преко 90% за други низ триплета је *SRoBERTa-F*, што је и очекивано јер су ова два модела обучавана на хрватским текстовима.

| редни број | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Andrija/SRoBERTa-base | Andrija/SRoBERTa-L | Andrija/SRoBERTa-XL | Andrija/SRoBERTa-F | classla/bcms-bertic | classla/xlm-r-bertic | xlm-roberta-base | xlm-roberta-large | jerteh/jerteh-81 | jerteh/jerteh-355 |
| $X_1$-$Y_1$ | 0,43 | 0,63 | 0,66 | 0,70 | / | 0,43 | 0,46 | 0,51 | 0,70 | **0,75** |
| $X_1$-$Y_2$ | 0,43 | 0,62 | 0,64 | 0,69 | / | 0,42 | 0,46 | 0,50 | 0,69 | **0,73** |
| $X_1$-$Y_3$ | 0,37 | 0,56 | 0,59 | 0,63 | / | 0,34 | 0,38 | 0,43 | 0,66 | **0,72** |
| $X_1$-$Y_4$ | 0,36 | 0,55 | 0,64 | **0,68** | / | 0,34 | 0,38 | 0,42 | 0,58 | 0,63 |
| $X_2$-$Y_1$ | 0,36 | 0,47 | 0,51 | 0,54 | / | 0,47 | 0,50 | 0,55 | 0,57 | **0,60** |
| $X_2$-$Y_2$ | 0,37 | 0,48 | 0,51 | 0,54 | / | 0,46 | 0,50 | 0,54 | 0,56 | **0,59** |
| $X_2$-$Y_3$ | 0,31 | 0,41 | 0,45 | 0,48 | / | 0,42 | 0,45 | 0,50 | 0,50 | **0,54** |
| $X_2$-$Y_4$ | 0,31 | 0,42 | 0,47 | **0,51** | / | 0,42 | 0,46 | **0,50** | 0,47 | **0,51** |
| $X_3$-$Y_1$ | 0,37 | 0,49 | 0,52 | 0,54 | / | 0,48 | 0,50 | 0,55 | 0,57 | **0,60** |
| $X_3$-$Y_2$ | 0,37 | 0,48 | 0,51 | 0,54 | / | 0,46 | 0,50 | 0,54 | 0,57 | **0,59** |
| $X_3$-$Y_3$ | 0,30 | 0,41 | 0,44 | 0,47 | / | 0,41 | 0,45 | 0,49 | 0,50 | **0,54** |
| $X_3$-$Y_4$ | 0,31 | 0,42 | 0,47 | **0,51** | / | 0,41 | 0,46 | **0,50** | 0,47 | **0,50** |
| $X_4$-$Y_1$ | 0,42 | 0,60 | 0,63 | 0,67 | / | 0,43 | 0,47 | 0,51 | 0,73 | **0,78** |
| $X_4$-$Y_2$ | 0,41 | 0,58 | 0,61 | 0,65 | / | 0,41 | 0,45 | 0,49 | 0,71 | **0,75** |
| $X_4$-$Y_3$ | 0,35 | 0,53 | 0,55 | 0,60 | / | 0,33 | 0,38 | 0,42 | 0,69 | **0,76** |
| $X_4$-$Y_4$ | 0,34 | 0,50 | 0,58 | 0,62 | / | 0,33 | 0,37 | 0,41 | 0,62 | **0,66** |
| просек | 0,36 | 0,51 | 0,55 | 0,59 | / | 0,41 | 0,45 | 0,49 | 0,60 | **0,64** |

**Табела 2.** Тачност модела у погађању токена (из три покушаја) на задатку моделовања маскираног језика над шеснаест припремљених маскираних текстова и њихова просечна тачност. Текстови су обележени (на почетку сваког реда) јединственим комбинацијама ознака токенизатора $X$ и извора $Y$. У сваком реду најбољи резултат ($\pm 1\%$) обележен је подебљањем.

Резултати који су модели остварили на низводним задацима $T_3$ (обележавање врсте речи) и $T_4$ (препознавање именованих ентитета) приказани су у виду $F_1$-мера у табели 4.

| редни број | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Andrija/SRoBERTa-base | Andrija/SRoBERTa-L | Andrija/SRoBERTa-XL | Andrija/SRoBERTa-F | classla/bcms-bertic | classla/xlm-r-bertic | xlm-roberta-base | xlm-roberta-large | jerteh/jerteh-81 | jerteh/jerteh-355 |
| $Y_1$-$Y_2$ | 0,93 | **0,95** | **0,96** | **0,96** | 0,92 | 0,76 | 0,90 | 0,87 | **0,95** | **0,95** |
| $Y_3$-$Y_4$ | 0,83 | 0,89 | **0,92** | **0,91** | 0,79 | 0,66 | 0,78 | 0,71 | 0,89 | 0,83 |
| просек | 0,88 | 0,92 | **0,94** | **0,93** | 0,85 | 0,71 | 0,84 | 0,79 | 0,92 | 0,89 |

**Табела 3.** Резултати одабраних модела на задатку препознавања реченица са истим значењем у триплетима екстрахованим из превода Достојевског ($Y_1$-$Y_2$), Верна ($Y_3$-$Y_4$), као и у просеку. У сваком реду најбољи резултат (±1%) обележен је подебљањем.

| р.б. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Andrija/SRoBERTa-base | Andrija/SRoBERTa-L | Andrija/SRoBERTa-XL | Andrija/SRoBERTa-F | classla/bcms-bertic | classla/xlm-r-bertic | xlm-roberta-base | xlm-roberta-large | jerteh/jerteh-81 | jerteh/jerteh-355 |
| $T_3$ | 0,974 | 0,980 | 0,982 | 0,982 | **0,986** | **0,987** | 0,984 | **0,986** | 0,985 | **0,986** |
| $T_4$ | 0,908 | 0,922 | 0,929 | 0,935 | **0,942** | **0,942** | 0,933 | 0,935 | 0,928 | 0,928 |

**Табела 4.** $F_1$-мера коју су модели остварили на задацима $T_3$ (обележавање врсте речи) и $T_4$ (препознавање именованих ентитета). У сваком реду најбољи резултат (±0.1%) обележен је подебљањем.

Из резултата приказаних у табели 4 види се да на задатку $T_3$ (обележавање врстом речи) девет од десет модела остварује јако добре резултате (преко 98%), при чему се резултати које остварују четири најбоља модела (*classla/bcms-bertic*, *classla/xlm-r-bertic*, *xlm-roberta-large* и *jerteh/jerteh-355*) разликују за мање од 0,02%, указујући да се модели полако приближавају горњим границама перформанси када је у питању овај задатак.

Када су у питању резултати остварени на последњем задатку $T_4$ (препознавање именованих ентитета), најбоље перформансе приказали су модели *classla/bcms-bertic* и *classla/xlm-r-bertic*, при чему су разлике у $F_1$-мерама нешто више него на задатку $T_3$ ($\sim 4\%$ за задатак $T_4$ у односу на $\sim 1\%$ за задатак $T_3$), али и даље знатно ниже него што су разлике на узводним задацима (чак $\sim 28\%$ за задатак $T_1$).

У наредном одељку биће разматрани резултати које су модели остварили, као и разлози који су довели до тих резултата, са циљем да се установе најповољнији услови за обучавање модела за српски језик у будућности.

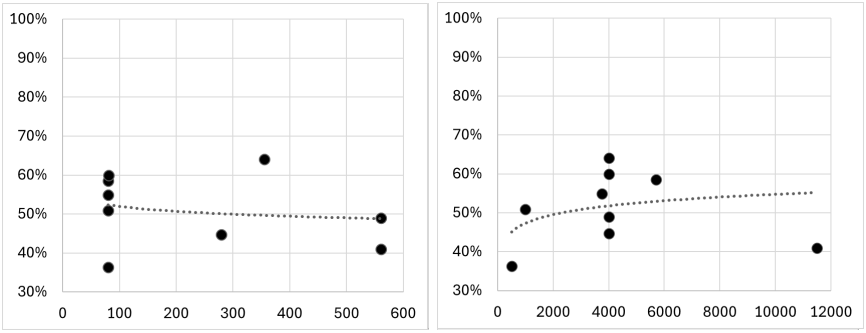
## 5. Дискусија

Претходно приказани резултати евалуације модела (табеле 2–4), показују да не постоји један модел, или група модела, који су најбољи у општем случају, већ су се различити модели показали као бољи (или лошији) на различитим задацима. У наставку, сваки од задатака биће посматран појединачно, пре свега у светлу односа остварених перформанси према величини модела, величини скупова за њихово обучавање и квалитету тих скупова.

### 5.1 Моделовање маскираног језика

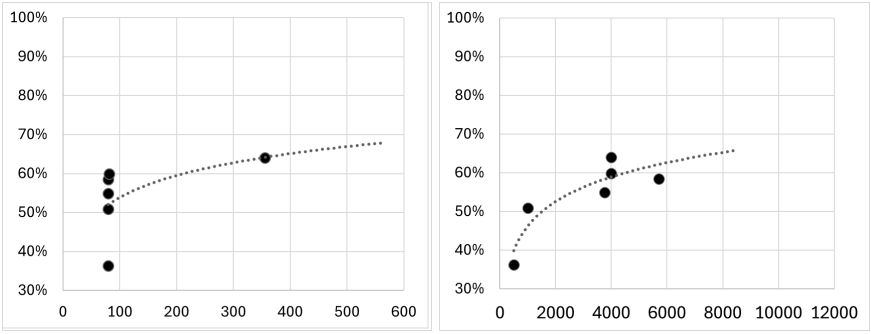
Резултати евалуације модела на задатку моделовања маскираног језика (табела 2) показују убедљиву предност модела *jerteh/jerteh-355*, са тим да добре резултате остварује и *jerteh/jerteh-81*, који је други најбољи модел за овај задатак када се посматрају просечне вредности. Пошто ова два модела користе исти скуп података за обучавање, то указује да би управо овај скуп могао бити разлог добрих резултата.

Просечна тачност модела на задатку $T_1$ према њиховој величини, односно према величини скупа који је коришћен за њихово обучавање приказан је на слици 1. На први поглед, приметни су неки истакнути




**Слика 1.** Тачност модела на задатку моделовања маскираног језика према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

изузеци, пре свега модели засновани на *XLM-R* архитектури, који остварују неке од најлошијих резултата на овом задатку. Уколико се њихови резултати уклоне, појављују се нови трендови (слика 2). Дакле, када посматрамо само *RoBERTa* моделе, чини се да већи модел (не баш убедљиво) и већи скуп за његово обучавање (врло убедљиво) повољно утичу на перформансе модела.

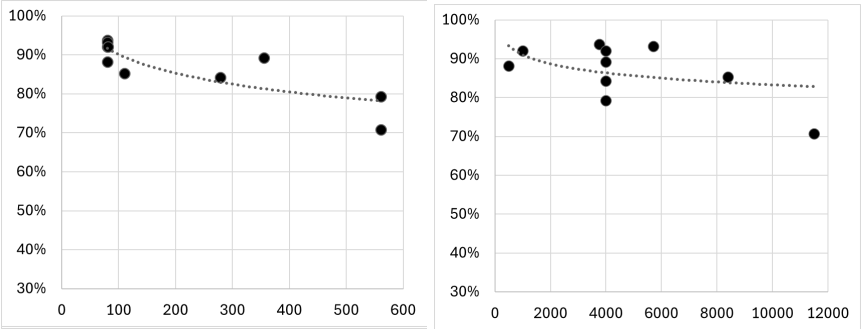



**Слика 2.** Тачност модела на задатку моделовања маскираног језика према величини тих модела (лево), као и према величини скупова за обучавање модела (десно), при чему су уклоњени резултати модела заснованих на *XLM-R* архитектури. Приказана крива тренда одговара логаритамској функцији.

### 5.2 Израчунавање сличности између реченица

Приликом евалуације задатка $T_2$ установљено је да најбоље резултате остварују модели *SRoBERTa-L*, *SRoBERTa-XL*, *SRoBERTa-F*, *jerteh-81* и *jerteh-355* када је у питању препознавање сличних реченица на српском језику, а *SRoBERTa-XL* и *SRoBERTa-F* када је у питању препознавање сличних реченица између српског и хрватског језика (табела 2). Ово указује да је кључ за добро угњежђивање реченица претходно обучавање модела за језике који се испитују.

Дакле, за угњежђивање реченица на српском језику најбољи су модели претходно обучени на довољно великом скупу реченица српског језика, али ако се обрађују и реченице на хрватском, модели који су обучавани и на српском и на хрватском имају предност. Са друге стране, модели засновани на *XLM-R* архитектури који су унапред обучени на сто светских језика поново показују најлошије резултате, вероватно због великог шума који разнолики скуп за обучавање производи.




**Слика 3.** Тачност модела на задатку угнежђивања према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

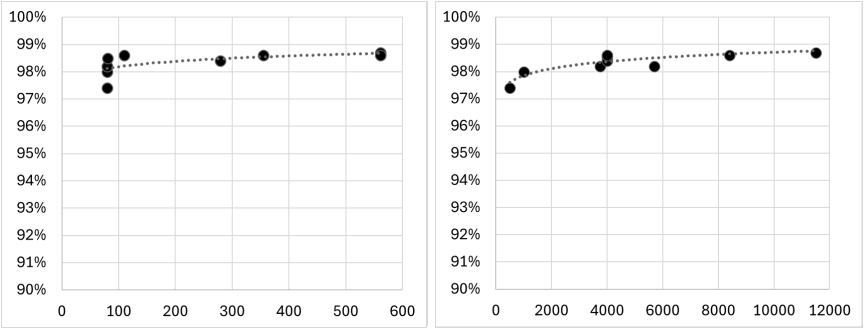
На слици 3 приказан је утицај величине модела и скупова за обучавање на перформансе на овом задатку, а линије тренда указују да се са повећањем и модела и скупа за обучавање перформансе смањују. Утицај величине скупа може донекле бити приписан претходно описаном феномену који погађа *XLM-R* архитектуру. Ипак, када је у питању утицај величине модела, постоје додатни индикатори да су мањи модели

бољи за овај задатак када је у питању обрада вишејезичних текстова. Тако *jerteh/jerteh-81* надмашује *jerteh/jerteh-355* на задатку утврђивања сличности између реченица на српском и хрватском језику. Разлог може бити то што је мањи модел, услед недостатка у величини, слабије прилагођен српском језику (*model underfit*), али зато има предност приликом генерализације.

### 5.3 Низводни задаци

За разлику од евалуације на узводним задацима, на низводним задацима резултати које постижу модели много су сличнији. На задатку обележавања врстом речи ($T_3$) готово сви модели остварују добре резултате (табела 4), укључујући и оне засноване на *XLM-R* архитектури. Штавише, *classla/xlm-r-bertic* и *xlm-roberta-large* два су од четири модела који остварују најбоље резултате (друга два модела су *classla/bcms-bertic* и *jerteh/jerteh-355*).

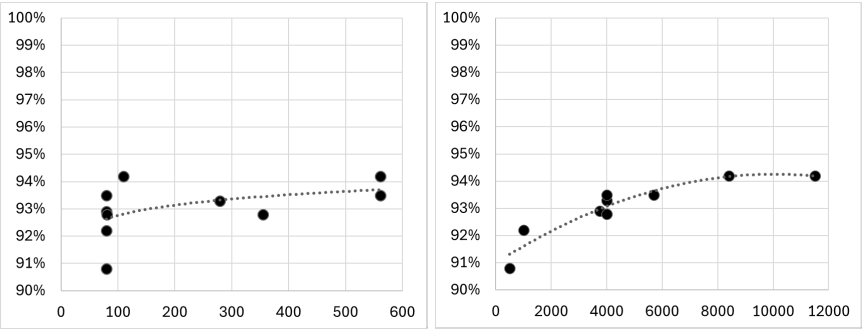
Заједничко за ова четири модела је да су или највећи модели или модели који су обучавани на највећим скуповима података. Позитивна корелација између перформанси и величине модела, као и између перформанси и величине скупова за обучавање уочава се и на слици 4.




**Слика 4.** Перформансе модела на задатку обележавања врстом речи према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

Корелација између величине скупа за обучавање још је очигледнија у случају препознавања именованих ентитета (слика 5). Најбоље резултате

на овом задатку ($T_4$) остварила су два модела са највећим скуповима за обучавање, док је поново приметна и успешност *XLM-R* модела, као и код претходног задатака. Величина модела такође показује тек незнатну позитивну корелацију.




**Слика 5.** Перформансе модела на задатку препознавања именованих ентитета према величини тих модела (лево), као и према величини скупова за обучавање модела (десно). Приказана крива тренда одговара логаритамској функцији.

### 5.4 Закључак

Када је у питању моделирање маскираног језика, чини се да развој нових модела за српски језик иде у правом смеру. Модел *jerteh/jerteh-355* остварује убедљиво најбоље резултате, барем када је у питању рад са висококвалитетним текстовима, чак и када су маскирани токенизаторима других модела (табела 1). Премда повећање скупа података за обучавање повољно утиче на перформансе модела (слика 2), не треба занемарити ни квалитет скупа јер модели *jerteh/jerteh-355* и *jerteh/jerteh-81* надмашују моделе *Andrija/SRoBERTa-F* и *Andrija/SRoBERTa-XL* обучене на већим скуповима, указујући да вебкорпуси можда нису увек довољни за обучавање добрих модела. Овај закључак је у складу са закључком из једног другог недавног истраживања (Li et al. 2023); ипак, ново истраживање би требало да у скуп за евалуацију уврсти и нелитерарне изворе, како би се добила свеобухватнија слика стања.

На задатку израчунавања сличности између реченица, тј. угнежђивања реченица, истакли су се модели као што су *Andrija/SRoBERTa-F* и *Andrija/SRoBERTa-XL*, за којима сасвим мало заостају *Andrija/SRoBERTa-L*, *jerteh/jerteh-81* и *jerteh/jerteh-355*, када су у питању реченице на српском језику (табела 3). Оно што издваја ове моделе је да су мањи у односу на друге моделе и да су обучавани на већем скупу, па изгледа да је за овај задатак кључна генерализација, дакле, већи скупови података (или можда мањи модели). Такође, када је у питању обрада реченица ширег језичког спектра (нпр. јужнословенски језици) било би потребно да се реченице из комплетног спектра укључе у скуп за обучавање или, боље, да се речник прилагоди за мапирање ширег спектра токена, па самим тим и за правилну векторизацију ових реченица. Ново истраживање на ову тему требало би да истражи и модернији начин векторизације реченица, на пример, коришћењем архитектуре трансформера реченица (*sentence transformers*) (Reimers and Gurevych 2019).

У случају оба *узводна задатка*, тачност коју остварују модели засновани на *XLM-R* архитектури много је нижа у односу на тачност модела заснованих на *RoBERTa* архитектури. У случају првог задатка ($T_1$) то се може објаснити њиховим већим речником токена (па је самим тим и избор одговарајућег токена тежи). Ипак, такво објашњење не би било адекватно и за други задатак ($T_2$). Са друге стране модели засновани на *XLM-R* архитектури показали су се као најбољи (уз малу маргину) на низводним задацима, пре свега на задатку препознавања именованих ентитета ($T_4$). Чини се да је за успешно решавање овог задатка најбоље да се модел током обучавања сусретне са најразличитијим бројем токена, али побољшања доноси и додатно обучавање на српским текстовима. Изгледа да би за унапређење перформанси тренутно било оптимално дообучавање *XLM-RoBERTa-large* модела коришћењем што већег и квалитетнијег скупа текстова на српском језику. Када је у питању обележавање врстом речи, изгледа да би било који нови модел био адекватан за решавање овог задатка.

## Захвалница

Најважније скупове података за обучавање модела *GPT2-orao*, *GPT2-vrabac*, *jerteh-81* и *jerteh-355* обезбедило је Друштво за језичке ресурсе и технологије.[12]

12. JeRTeh

Рачунарске ресурсе за обучавање модела *GPT2-orao* и *GPT2-vrabac* обезбедила је Национална платформа за вештачку интелигенцију Србије.

Рачунарске ресурсе за обучавање модела *jerteh-81* и *jerteh-355* обезбедио је Рударско-геолошки факултет Универзитета у Београду.

Истраживање је спроведено уз подршку Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

Хвала!

## Литература

Biemann, Chris, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. "The Leipzig corpora collection-monolingual corpora of standard size." *Proceedings of corpus linguistic* 2007.

Bogdanović, Miloš, Jelena Kocić, and Leonid Stoimenov. 2024. "SRBerta-A Transformer Language Model for Serbian Cyrillic Legal Texts." *Information* 15 (2). ISSN: 2078-2489. https://doi.org/10.3390/info15020074.

Clark, Kevin, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555.*

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116.*

Cvejić, Andrija. 2022. "Prepoznavanje imenovanih entiteta u srpskom jeziku pomoću transformer arhitekture." *Zbornik radova Fakulteta tehničkih nauka u Novom Sadu* 37 (02): 310–315.

Cvetanović, Aleksa, and Predrag Tadić. 2023. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." In *2023 31st Telecommunications Forum (TELFOR),* 1–4. IEEE.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Dobreva, Jovana, Tashko Pavlov, Kostadin Mishev, Monika Simjanoska, Stojancho Tudzarski, Dimitar Trajanov, and Ljupcho Kocarev. 2022. "MACEDONIZER-The Macedonian Transformer Language Model." In *International Conference on ICT Innovations,* 51–62. Springer.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. "DEBERTA: Decoding-Enhanced BERT with Disentangled Attention." In *International Conference on Learning Representations.*

Jiang, Albert Q, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. "Mistral 7B." *arXiv preprint arXiv:2310.06825.*

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 7871–7880.

Li, Yuanzhi, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. "Textbooks are all you need ii: phi-1.5 technical report." *arXiv preprint arXiv:2309.05463.*

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Roberta: A robustly optimized BERT pretraining approach." *arXiv preprint arXiv:1907.11692.*

Ljubešić, Nikola, and Filip Klubička. 2014. "{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian." In *Proceedings of the 9th web as corpus workshop (WaC-9),* 29–35.

Ljubešić, Nikola, and Davor Lauc. 2021. "BERTić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian." *arXiv preprint arXiv:2104.09243.*

Ljubešić, Nikola, Vit Suchomel, Peter Rupnik, Taja Kuzman, and Rik van Noord. 2024. *Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining.* Submitted for review.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. "Improving language understanding by generative pre-training."

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. "Language models are unsupervised multitask learners."

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21 (1): 5485–5551.

Rajpurkar, Pranav, Robin Jia, and Percy Liang. 2018. "Know what you don't know: Unanswerable questions for SQuAD." *arXiv preprint arXiv:1806.03822.*

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 3982–3992.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian ner&beyond: The archaic and the modern intertwinned." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260.

Škorić, Mihailo. 2023. "Композитне псеудограматике засноване на паралелним језичким моделима српског језика." Докторска дисертација. PhD diss., Универзитет у Београду.

Škorić, Mihailo, Miloš Utvić, and Ranka Stanković. 2023. "Transformer-Based Composite Language Models for Text Evaluation and Classification." *Mathematics* 11 (22): 4660.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3954–3962.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary. 2019. "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures." In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7).* Leibniz-Institut für Deutsche Sprache.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. "Alpaca: A strong, replicable instruction-following model." *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3 (6): 7.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *Advances in neural information processing systems* 30.

Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. 2008. "Tour du monde through the dictionaries." In *Actes du 27eme Colloque International sur le Lexique et la Gammaire,* 249–256.

Wasserscheidt, Philipp. 2023. *Serbian Web Corpus PDRS 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1752.

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data" [in English]. In *Proceedings of the 12th Language Resources and Evaluation Conference,* 4003–4012. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. https://www.aclweb.org/anthology/2020.lrec-1.494.

# Изградња речника фудбалске терминологије вођена подацима и OntoLex моделом

УДК 811.163.41'322.2

**САЖЕТАК:** Рад истражује иновативни приступ развоју речника фудбалске терминологије применом повезаних података (енгл. *Linked Data*). Традиционални приступи изради речника често укључују секвенцијално писање речничких чланака и ручну анализу текстуалних корпуса, што може бити споро и захтевно. Наш приступ користи технике рачунарске лингвистике и аутоматизацију како би се убрзао и побољшао процес креирања речника. Дигитални речник фудбалске терминологије који смо развили намењен је и људима и софтверским апликацијама, пружајући корисницима брз приступ, могућност сталног ажурирања и интеграцију са другим дигиталним ресурсима. Овај рад позива се на српски део српско-шпанског речника фудбалских термина, који се развија у оквиру докторске дисертације Јелене Лазаревић.

**КЉУЧНЕ РЕЧИ:** речник, фудбал, корпус, повезани подаци, рачунарска лингвистика.

**РАД ПРИМЉЕН:** 10. јун 2024.
**РАД ПРИХВАЋЕН:** 27. јун 2024.

Јелена Лазаревић
jelazarevic1@gmail.com
ORCID: 0009-0004-4481-2729
*докторанд*
*Универзитет у Београду*
*Филолошки факултет*
*Београд, Србија*

Оливера Китановић
olivera.kitanovic@rgf.bg.ac.rs
ORCID: 0000-0002-7571-2729
*Универзитет у Београду*
*Рударско-геолошки факултет*
*Београд, Србија*

## 1. Увод

Интеграција језичких ресурса једна је од главних тема у савременим истраживањима у вези са језичким технологијама. У порасту је примењивање приступа вођеног подацима (енгл. *data-driven approach*) ради побољшаних могућности аутоматизације обраде података. Приступ вођен подацима је методологија која се ослања на анализу и тумачење података као основе за доношење одлука и решавање

проблема. Уместо интуиције, искуства или претходних теоријских знања, приступ вођен подацима користи конкретне, објективне и мерљиве информације добијене из података. Овај приступ је изузетно популаран у научним дисциплинама, попут лингвистике, медицине, различитих информатичких технологија (Kitanović et al. 2021). У овом раду представићемо истраживање могућности убрзања и побољшања традиционалног приступа изради речника кроз пример развоја двојезичног речника фудбалских термина. Традиционални приступ креирања речничких чланака, тачније писања једног по једног, најчешће у алфабетском или азбучном редоследу, ослања се на анализу грађе у виду корпуса или ексцерпираних примера уз коришћење референтних речника за сваки појединачни термин. Помоћу технологија рачунарске лингвистике процес се убрзава и аутоматизује. Наиме, чланци се не обрађују секвенцијално, већ је примењен приступ вођен подацима и паралелном обрадом већег скупа података одједном.

Берг и Оландер (Bergh and Ohlander 2019) уочили су да је у XX веку речи фудбалске терминологије постале присутније, а неки термини су постали уобичајени у говору навијача. Језик фудбала је у сталној промени, прилагођава се догађајима у самој игри и око ње. Због свог значаја и велике медијске присутности, фудбал као „народна игра“ утиче на то да енглески речници опште намене укључују фудбалске термине као део општег језика.

Дигитални речници имају бројне предности у односу на традиционалне, штампане речнике. Фуертес-Оливера и Тарп (Fuertes-Olivera and Tarp 2014) сматрају да је идеално решење за изазове новог доба управо онлајн – дигитално издање речника. Прва и кључна предност речника доступног на интернету је брзина претраге речи или фраза, будући да се у штампаним речницима странице морају ручно прелиставати. Друга важна предност јесте могућност сталног ажурирања, насупрот штампаним речницима који могу застарети током времена. Дигитални речници омогућавају прилагођавање врсте фонта, величине слова и других параметара како би задовољили потребе корисника, а могу садржати и додатне информације, попут детаљно објашњене етимологије и изговора речи. Доступност путем мобилних апликација или рачунарских платформи омогућава корисницима приступ речнику било када и било где. Интерактивни елементи као што су хиперлинкови и мултимедијални садржаји чине искуство учења језика динамичнијим. Такође, дигитални речници често садрже директну претрагу унутар текста речничког чланка, што је корисно за

брзо проналажење значења или информација о одређеној речи док се чита неки текст.

Интегрисање речника у дигиталне платформе за учење језика омогућава праћење напредовања у знању језика кроз прављење личних листа речи и чувања бележака, као креирање картица за подсећање и других алата којима се олакшава учење језика. Интегрисање речника са корпусима текстова је веома значајно и стога што корисници могу да уче из конкретних примера њихове употребе – било да су у питању појединачне речи, фразе или конструкције.

Насупрот светској пракси, у Србији штампани речници и даље имају већи степен употребе односу на дигиталне, услед осећаја аутентичности искуства или навика употребе. У савременом добу, дигитални речници постали су неопходан алат за оне који уче и користе стране језике. Не треба заборавити да је из дигиталног облика речника могуће припремити за штампу више различитих облика и обима речника, у различитим форматима, са разним нивоима детаљности и већим бројем речничких чланака.

Речник фудбалске терминологије који представљамо у раду је дигитални речник намењен и људима и машинама; тачније, софтверским апликацијама и део је ширег истраживања и изградње првог дигиталног фудбалског речника на српском језику, са преводима на шпански језик. У контексту истраживања насталог као део докторске дисертације Јелене Лазаревић под називом: „Језичке одлике дискурса нових медија о фудбалу: контрастивна анализа на корпусу српског и шпанског језика.“ Нагласак је на спрези техника рачунарске лингвистике, као и примени технологија семантичке мреже, како би се задовољиле потребе појединца и корисника, али и машине, односно софтверских алата као алтернативних корисника речника.

Речник је намењен спортским радницима: спортистима, тренерима, судијама, професорима, студентима, новинарима, као и свима онима који желе да овладају прецизнијом интерпретацијом фудбалских термина. Овај рад се посебно фокусира на српски део двојезичког српско-шпанског речника фудбалских термина.

У одељку 2. овог рада биће сагледана два начина издвајања информација потребних за изградњу речника: аутоматска ексцерпција из корпуса *срФудКо* кандидата за термине, фреквенција и примера употребе (Krstev et al. 2015; Ivanović et al. 2022) и екстракција информација из постојећих речника и глосара доступних онлајн (Kitanović et al. 2021).

## 2. Припрема података за изградњу речника

Корпус *срФудКо* (Lazarević et al. 2023) изграђен је од новинских чланака о фудбалу на српском језику који су прикупљени са пет српских дигиталних новинских сајтова: *Б92, Блиц, Мондо, Политика* и *Спорт Клуб*. Чланци су аутоматски преузети коришћењем различитих техника за прикупљање података са интернета, након чега је уследила дедупликација текстова, елиминисање кратких чланака (краћих од 3.000 карактера), реченица на другим језицима и табела које садрже само нумеричке резултате. Након филтрирања чланака и чишћења текста, креиран је корпус *срФудКо* који садржи 10.100.553 токена, од којих су 8.618.426 речи, док остатак чине знакови интерпункције. Садржај корпуса је анотиран врстама речи и лематизован коришћењем тагера за српски језик *SrpKor4Tagging-TreeTagger* (Stanković et al. 2020; Stanković, Škorić, and Šandrih Todorović 2022).

Корпус *срФудКо* је затим коришћен за ексцерпцију информација. Први резултат аутоматске ексцерпције из корпуса је листа кандидата речи, праћена фреквенцијама појављивања у корпусу. Затим су издвојени илустративни примери употребе појединих термина. У првој фази су екстраховани, евалуирани и обележени једночлани термини, а уследила је примењена екстракција вишечланих термина (Krstev et al. 2015) најчешћих синтаксичких образаца, заснована на морфолошким речницима српског језика и локалним граматика (Krstev 2008). Осим фреквенција, за једночлане термине је рачуната и кључност[1] (енгл. *keyness*) термина, где су поређене фреквенције у корпусу са фреквенцијама у корпусу савременог српског језика *СрпКор2013* (Утвић 2011). Када су у питању вишечлани термини коришћене су комплексније асоцијативне мере *T-Score, CValue*, и друге (Vu, Aw, and Zhang 2008). Детаљи ове фазе развоја речника су описани у раду: *Football terminology: compilation and transformation into OntoLex-Lemon resource* (Lazarević et al. 2023).

Када је реч о екстракцији информација из постојећих речника и глосара доступних на интернету, преузети су речници различитог типа и нивоа детаљности. Уместо да се за сваки чланак појединачно користи речник, планирано је укрштање, усклађивање и обрада свих извора како би се олакшала и убрзала изградња савременог фудбалског речника богатог информацијама.

Интеграција података се ослањала на више извора, али главну улогу су имали: 1) четворојезични речник, српско-енглеско-

1. https://www.sketchengine.eu/documentation/simple-maths/

француско-шпанска листа термина (Mihajlović 2003) и 2) *Kicktionary*[2] (Schmidt 2009) заснован на концепту семантике оквира по узору на *FrameNet*[3] (Fillmore et al. 2003). Семантика оквира је теорија значења која се фокусира на концептуалне структуре (оквире) које представљају позадину потребну за разумевање речи и израза у различитим контекстима. Ови оквири укључују повезане елементе и односе који се активирају приликом употребе неке речи или израза, омогућавајући боље разумевање њиховог значења и контекста. Оквири и лексичке јединице на енглеском су укрштене са енглеском страном четворојезичног речника како би се обогатила српска страна речника семантичким оквирима.

С обзиром на чињеницу да ниједан српски речник фудбалске терминологије нема описне дефиниције појмова, а имајући у виду напредак машинског превођења, зарад брже израде речника, који је предмет истраживања, користили смо глосаре фудбалске терминологије на енглеском и шпанском језику и то полуаутоматском обрадом. Осим речника *Kicktionary*, који такође садржи одређене дефиниције, коришћени су и речници доступни у тренутку обраде. Неопходно је напоменути да су неки од речника у међувремену постали недоступни на раније познатим адресама:

– Глосар *Field*[4] са 900 најчешће коришћених термина организован је кроз илустроване сценарије. Изворни језик је португалски а садржи еквиваленте на енглеском и шпанском, заједно са примерима употребе. Поглед на речник омогућен је кроз листу фудбалских термина и кроз сценарије. Хомонимија и полисемија је имплементирана кроз одвојене речничке чланке и генерално је структура инспирисана концептима семантике оквира. Анализа полисемичних речи и колокација у корпусу укључила је, поред осталих фаза, компилацију трију упоредивих корпуса за шпански, португалски и енглески.

2. http://www.kicktionary.de/ вишејезички лексикон који обухвата појмове и њихове дефиниције везане за фудбалске утакмице, тактике, играче, и опрему, са циљем да олакша учење и разумевање фудбалске терминологије кроз визуелизације и објашњења.

3. FrameNet је лексичка база података која документује семантичке оквире и како они повезују речи са њиховим значењима у различитим контекстима.

4. раније доступан на http://dicionariofield.com.br/, видети https://www.facebook.com/dicionariofield

– Фудбалски глосар *UEFA*[5] који смо користили има укупно 8.086 речничких чланака, од тога на немачком 1.893, на енглеском 2.306 и на шпанском 1.887.
– Фудбалски глосар *UsingEnglish*[6] као део речника који садржи различите домене има свега 74 речничка чланка за фреквентне одреднице.
– Деветојезични речник[7] упарених фудбалских термина има 77 термина са преводним еквивалентима на енглеском, немачком, француском, шпанском, италијанском, португалском, пољском, руском и украјинском.

Речник *ФудЛе* настао на основу докторске дисертације Јелене Лазаревић садржи сада 2.648 српско-шпанских преводних еквивалената којима су придружене врсте речи и фреквенције у корпусу *срФудКо*, при чему 93 речничка чланка имају и придружену дефиницију. У српском делу речника једночланих термина има 990, двочланих има 966, са три компоненте 369, са четири 200, а са пет и више има 123. Повезивање са семантичким оквирима речника *Kicktionary* је урађено за 142 лексичке јединице. Након екстракције терминолошких фраза из корпуса *срФудКо*, ручне евалуације и обележавања доменских маркера: да ли је у питању спортски термин или је у питању термин карактеристичан за фудбалски домен, добијена је листа од 1.943 термина којима су придружени код синтаксичке групе, фреквенција и мере асоцијације.

## 3. Компилација и трансформација речника у OntoLex модел

Употреба модела лексикона за онтологије *OntoLex-Lemon*,[8] краће *OntoLex-a* (*LExicon Model for ONtologies*) (McCrae et al. 2017), у порасту је када је у питању имплементација лексичких ресурса као скупа повезаних података на интернету. Одреднице, без обира на то да ли се односе на једночлане или вишечлане термине из домена фудбала, заједно са резултатима екстракције из корпуса *срФудКо* (фреквенције и

5. раније доступан на https://www.uefa.com/insideuefa/dictionary/
6. https://www.usingenglish.com/glossary/football-vocabulary/
7. https://www.uefa.com/MultimediaFiles/Download/competitions/General/01/80/75/52/1807552_DOWNLOAD.pdf
8. https://www.w3.org/2016/05/ontolex/,https://jogracia.github.io/ontolex-lexicog/

примери), обогаћени информацијама преузетим и обрађеним из речника и глосара, представљени су коришћењем модела *OntoLex*.

Поред превода, речнички чланак указује на врсту речи појединачне одреднице, и потенцијално граматичка својства. Додатне информације о термину у виду дефиниције значења, везе ка концептима у другим ресурсима, попут база знања, онтологија или лексичких база значајне су кориснику ради бољег разумевања конкретног термина. Контекст употребе кроз изабране примере екстраховане из корпуса уз фреквенције појављивања у корпусу треба да пружи додатне информације и употпуни терминолошку слику.

Треба напоменути да су прве употребе *Lemon*, а потом и *OntoLex-Lemon* модела за српски језик повезане са прављењем лексичке базе Лексимирка[9] (Stanković et al. 2018) и превођење система електронских речника за српски (Krstev 2008) у релациону базу (Рујевић 2022). У раду се користи модел *OntoLex* за моделирање RDF[10] (енгл. *Resource Description Framework*), односно представљања лингвистичких описа лексичких јединица из генерисаног речника уз проширење лексичког слоја, како би се обезбедило њихово повезивање и машинска разумљивост на интернету. *OntoLex* модули су уведени кроз примере имплементације, почевши са основним (*OntoLex*) модулом који дефинише лексичке јединице, облике речи и лексичка значења. Модул *vartrans*[11] има две важне класе: *Translation* и *TranslationSet*, где превод представља везу између двају лексичких значења, повезаних са лексичким јединицама на различитим језицима. *LexInfo*[12] онтологија типова, вредности и својстава користи се као каталог категорија података (нпр. за означавање граматичког рода, броја, врсте речи итд.).

Категорије превода су представљене упућивањем на екстерни каталог *OEG Translation Categories*.[13] Други често коришћени речници, као што је *Dublin Core, DCMI Metadata Terms*,[14] користе се за метаподатке о изворима, ауторству, верзији или лиценци. За моделирање следиле су се препоруке и приступи дати у смерницама *Guidelines for*

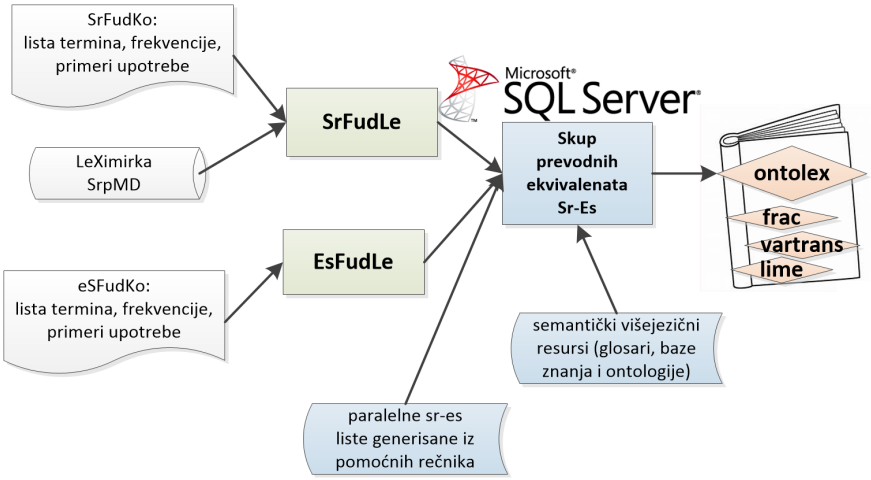
9. https://leximirka.jerteh.rs/
10. https://www.w3.org/RDF/
11. https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans
12. https://lexinfo.net/, https://github.com/ontolex/lexinfo
13. http://purl.org/net/translation-categories
14. http://purl.org/dc/elements/1.1/

*Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0)*[15] (Río Gracia et al. 2023), и низу смерница „Guidelines and best practices for LLOD“.

Према наведеним препорукама, двојезични речник се организује тако да су појединачни језици и одвојени RDF графови, који се потом повезују додатним графом који садржи везе између преводних еквивалената. Генерисани речник се састоји из четири графа, односно скупа података:

– Изворни лексикон, у овом случају српски,
– Циљни лексикони, у овом случају шпански,
– Скуп преводних еквивалената (енгл. *TranslationSet*).

Приступ коришћења ресурса припремљених у претходним фазама (леви и доњи део слике) и модели коришћени за трансформисање у RDF облик (десни део слике) илустровани на слици 1, тако да се на излазу добија српско-шпански двојезични речник *ФудЛе* чији су речнички чланци обогаћени бројним информацијама.




**Слика 1.** Графички приказ интеграције језичких података из различитих ресурса у речник *ФудЛе*.

15. https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

OntoLex-FrAC модул (Chiarcos et al. 2020) коришћен је да у речник фудбалске терминологије укључи информације о фреквенцијама појављивања у корпусима, примере употребе термина у контексту и колокацијама екстрахованим из корпуса *срФудКо*. Аутоматски екстраховани термини, који су затим ручно евалуирани и обележени етикетама за домен спорт или фудбал, допуњени су флективним облицима који се даље трансформишу у модел *OntoLex*. Даље су термини повезани са примерима из корпуса који су одабрани коришћењем GDEX[16] (Good Dictionary Examples) алгоритма(Kilgarriff et al. 2008).

Морфолошки речник вишечланих термина је направљен коришћењем алата LeXimir (Stanković et al. 2011), а затим је трансформисан апликацијом која прати *OntoLex* спецификације и објављене примере (Chiarcos, Fäth, and Ionov 2022). Граматичке информације, морфосинтаксичка својства облика речи дате су у складу са LexInfo вокабуларом. У наредним одељцима представљамо употребу основног модула модела *OntoLex* и модула за фреквенције *OntoLex-FrAC*, примере употребе и остале информације засноване на корпусу (Chiarcos et al. 2022).

### 3.1 Језгро речника *ФудЛе*

Корпус *срФудКо* је латинични, стога је и речник *ФудЛе* написан латиничним писмом. Пример термина: „фудбалска утакмица“ је приказан у раду (Lazarević et al. 2023), док у овом раду дајемо пример приказа термина „жути картон“ у моделу *OntoLex*, који је проширен додатним информацијама о значењу и релацијама ка другим ресурсима и језицима. Жути картон се дефинише као упозорење које судија упућује играчу због неспортског понашања. Судија жутим картоном санкционише играча за фаул у случају када понашање играча не приказује тежину прекршаја који заслужује црвени картон, као што је туча или фаул са намером да повреди противника. Овај термин постоји и у српском морфолошком речнику SrpMD сложених речи (Krstev and Vitas 2009) у облику DELAC записа (Savary, Krstev, and Vitas 2007), а његов изворни запис је:

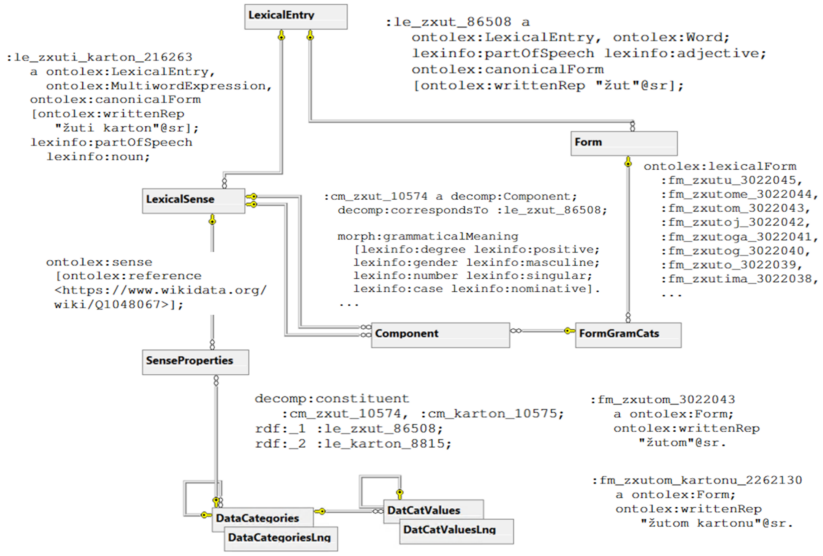
```
žuti(žut.A8:adms1g) karton(karton.N1:ms1q),
NC_AXN+DOM=Sport+Comp+Conc
```

16. https://www.sketchengine.eu/documentation/manual-for-gdex/

Коначни трансдуктор NC_AXN генерише флективне облике за морфолошке електронске речнике сложених речи, где NC означава именичку колокацију, а AXN приказује колокацију облика придев-именица, где се придев слаже са именицом по свом граматичком броју, роду, падежу и аниматности. За компоненте које трансдуктор укључује, захтевају се информације о леми (жути и картон), трансуктору (A8 и N1) за појединачне компонентне колокације и вредности за граматичка својства (adms1g и ms1q). Граматичка својства су следећа: a - позитиван степен, d - одређени (дуги) облик, m - граматички мушки род, s - једнина, 1 - номинатив, g - аниматност није од значаја, q - неживо. Већина граматичких својстава се лако пресликава у онтологију LexInfo, али опција када аниматност није од значаја није предвиђена. Имајући у виду да се термин: „жути картон“ може наћи у неколико терминолошких речника, коришћењем модела *OntoLex*, овој лекскикализованој колокацији додељујемо и њено специфично значење.

Део модела података базе података система *LeXimirka*, имплементиране у MS SQL Серверу (Stanković et al. 2018; Рујевић 2022), приказан је на слици 2 која приказује табеле за лексичке јединице (*LexicalEntry*) и флективне облике (*Form*), као и компоненте за вишечлане речи (*Component*). Једна од улога система LeXimirka је генерисање флективних облика и њихових граматичких информација, за једночлане и вишечлане термине речника *ФудЛе*. Граматичке информације су повезане са флективним облицима кроз категорије података (*DataCategories*) и њихове вредности (*DatCatValues*). Систем је опремљен метаподацима који се односе на повезане информације између категорија података у српским морфолошким речницима и речнику *LexInfo*. Једна лексичка јединица може имати више значења, сачувана у табели *LexicalSense*, која појединачне категорије прикупља преко функције *SenseProperties* из каталога *DataCategories*.

Запис на слици 2 приказује колокацију „жути картон“, где се назив лексичке јединице *le_zxuti_karton_216263* састоји од префикса le који потиче од LexicalEntry, самог термина, односно колокације: „жути картон“ и примарног кључа 216263 у табели LexicalEntry базе података LeXimirka. Када су у питању компоненте, слично се бира префикс cm да обележи записе који долазе из табеле *Component*, док префикс *fm* означава записе из табеле *Form*. Напоменимо да се за поједина слова мењају према кодирању Аурора (Vitas 1979) који је осмислио проф. др Душко Витас како би се обезбедило једнозначно ASCII кодирање које је




**Слика 2.** Дијаграм базе података у MS SQL Серверу са повезаним табелама и примером експортованих података у *OntoLex*.

погодно за називе објеката скупа података, док се литерали, тј. писани облици, кодирају коришћењем кодне шеме UTF-8.

За лексичку јединицу *:le_zxuti_karton_216263* видимо да припада класи *ontolex:LexicalEntry* и да је у питању вишечлани термин (*ontolex:MultiwordExpression*), коју наводимо у канонском облику, односно као лему (*ontolex:canonicalForm*). Затим се наводи писани облик леме, уз податак да је у питању именица. Када је значење у питању, упућује се на два екстерна ресурса. Први је база знања Википодаци (Wikidata) која има одговарајући ставку „жути картон“ чији је код Q1048067 и целокупна URL адреса: https://www.wikidata.org/wiki/Q1048067. Има преводне еквиваленте на 38 језика, па осим лексикализације у српском, такође има и шпански превод: “tarjeta amarilla”.

```
:le_zxuti_karton_216263 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "žuti karton"@sr];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>.
```

```
  decomp:constituent :cm_zxuti_10574, :cm_karton_10575;
    rdf:_1 :le_zxut_86508;
    rdf:_2 :le_karton_8815.

# komponente kanonskog oblika
:cm_zxuti_10574 a decomp:Component;
  decomp:correspondsTo :le_zxut_86508;

  morph:grammaticalMeaning [
    lexinfo:degree lexinfo:positive;
    lexinfo:gender lexinfo:masculine;
    lexinfo:number lexinfo:singular;
    lexinfo:case lexinfo:nominative].
  ...
```

Припадајући део на шпанском језику можемо видети у наставку:

```
:le_tarjeta_amarilla_10001 a ontolex:LexicalEntry,
   ontolex:MultiwordExpression,   ontolex:canonicalForm
   [ontolex:writtenRep "tarjeta amarilla"@es];
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:denotes <https://www.wikidata.org/wiki/Q1048067>;

  decomp:constituent :cm_tarjeta_1, :cm_amarilla_2;
    rdf:_1 :le_tarjeta_1;
    rdf:_2 :le_amarillo_2.
```

### 3.2 Дефинисање значења у речнику *ФудЛе*

Класа *ontolex:LexicalSense* дефинише лексичко значење лексичке јединице када се тумачи као да се односи на одговарајући онтолошки елемент. Лексичко значење представља реификацију или опредмећење упаривања лексичке јединице и онтолошке целине на коју се односи. Веза између лексичке јединице и онтолошког ентитета преко објекта *ontolex:LexicalSense* имплицира да се лексичка јединица може користити за упућивање на задати онтолошки ентитет.

Идеја прецизнијег повезивања лексикона *FrameNet* и *OntoLex* модела приказана је у (Robin, Kulkarni, and Buitelaar 2023). Код дефинисања лексичких јединица показано је да се могу користити за означавање одређеног онтолошког предиката, коришћењем својства: *ontolex:denotes*. Дакле, лексичка јединица означава предметну класу или онтолошки елемент. Уз то је могуће представити евокацију одређеног менталног концепта од задате лексичке јединице, која се не односи на класу са формалном интерпретацијом у неком моделу. Класа лексички концепт (*ontolex:LexicalConcept*) представља менталну апстракцију, концепт или јединицу мишљења која се може лексикализовати, тако да класа *ontolex:LexicalConcept* представља потклasу елемента *skos:Concept*.

Дефиниције се додају лексичком концепту као глоса и то коришћењем својства *skos:definition*.

Истражујући изградњу семантичког речника фудбала, ослањајући се на корпус из фудбалског домена, Wu i Li (Wu and Li 2017) користили су Word2Vec (Mikolov et al. 2013) ради обучавања векторског модела за речи у корпусу. Комбинујући статистички алгоритам TF-IDF и векторе речи, одређивали су концепте карактеристичне за фудбалски домен. За листу термина из области фудбала израчунали су растојања од речи у векторском моделу и издвојили три најближе речи, то јест, термина са вишом вредношћу косинуса као синониме за улазну реч за изградњу семантичког речника. Показали су да постојећи семантички речници, као што је WordNet, не одговарају потребама појединих домена зато што је семантичка грануларност речи мања. Дакле, потребно је направити семантички речник који за један појам повезује синониме, као и подређене и надређене појмове и дефиниције на различитим језицима. Тиме се добија семантички речник којим се имплементира семантичко претраживање текстова из фудбалског домена.

Креирали смо сопствени идентификатор концепта на основу прикупљених ресурса. Речник *UEFA*, са терминима и дефиницијама је узет као узор за дефинисање концепата. Конкретно, пример је *:ct_ yellow_ card_ 1*, где је ct скраћено од "concept", потом је узет енглески термин и додат идентификатор (ради једнозначности) и уместо размака се користи доња црта "_".

```
:conceptLexicon a ontolex:ConceptSet .
:le_zxuti_karton_216263 ontolex:evokes :ct_yellow_card_1.
:le_tarjeta_amarilla_10001 ontolex:evokes :ct_yellow_card_1.

:ct_yellow_card_1 a ontolex:LexicalConcept ;
  ontolex:LexicalisedSense :sn_zxuti_karton_216263, :sn_tarjeta_amarilla_10001;
  ontolex:isConceptOf  <https://www.wikidata.org/wiki/Q1048067> ;
  skos:definition "Disciplinska sankcija koju sudija izriče tokom utakmice
    protiv igrača zbog kršenja Pravila igre, posebno zbog lošeg ponašanja u
    sportu ili odlaganja ponovnog početka igre."@sr;
  skos:definición "Sanción disciplinaria impuesta durante un partido por el
    árbitro a un jugador por infracción de las Reglas de Juego, en particular
    por mala conducta deportiva o por retrasar la reanudación del juego."@es;

ontolex:isEvokedBy :le_zxuti_karton_216263, sn_tarjeta_amarilla_10001;
  skos:inScheme :FudLe .
```

Имајући у виду да се за ontolex:LexicalSense може везати глоса која представља специфичности коришћења употребом својства ontolex:usage. Наиме, интерпретација лексичке јединице у односу на значење дефинисано у датој онтологији је често прилагођено условима

употребе или прагматичним импликацијама, посебно због регистра, конотација или варијација у значењу. Ове нијансе значења могу се спецификовати коришћењем својства, што омогућава прикупљање информација о условима употребе и прагматичким импликацијама, под којима се лексички унос упућује на онтолошко значење.

Поменути услови употребе не уводе се уместо формално дефинисаног значења, већ допуњују одговарајуће значење додатним информацијама које описују употребу лексичке јединице. У случају речника *ФудЛе* додатне дефиниције прикупљене из различитих извора, екстраховане из корпуса или директно написане, уписиване су на овај начин.

```
:sn_zxuti_karton_216263 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html>;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_zxuti_karton_216263 ;
  ontolex:usage [ rdf:value "Žuti karton koji sudija vadi iz svog džepa da bi pokazao
    igraču koji je uradio nešto prilično loše, kao što je loš faul ili igranje rukom,
    ali ne tako loš kao prekršaj crvenog kartona kao što je tuča. Žuti karton često
    ide zajedno sa drugom kaznom kao što je slobodan udarac. Dva žuta kartona na jednom
    meču znače crveni karton, a dva žuta u određenom periodu suspenziju."@sr ].

:sn_tarjeta_amarilla_10001 a ontolex:LexicalSense ;
  ontolex:reference <http://www.kicktionary.de/LUs/Sanction/LU_1240.html> ;
  ontolex:isLexicalizedSenseOf :ct_yellow_card_1 ;
  ontolex:isSenseOf :le_tarjeta_amarilla_10001 ;
  ontolex:usage [ rdf:value "La tarjeta amarilla se muestra a un jugador por acciones
    como faltas fuertes o tocar el balón con la mano. Puede ir acompañada de sanciones
    como tiros libres. Dos tarjetas amarillas en un partido significan una roja, y dos
    en un período llevan a suspensión."@es ].
```

### 3.3 Имплементација фреквенција и примера у речнику *ФудЛе*

Документација за модул *OntoLex* и информације о фреквенцијама, примерима употребе у контексту и другим корпусним информацијама “Frequency, Attestations and Corpus data” (FrAC) налази се на линку (Chiarcos et al. 2022; Chiarcos et al. 2020). Модул FrAC је усмерен на допуну речника и других лингвистичких ресурса који садрже лексикографске податке моделом за репрезентацију статистике изведене из корпуса: информације о учесталости и појављивању, колокације, показивачима из лексичких ресурса на корпусе и друге текстуалне колекције: потврде употребе у корпусној грађи, обележавање корпуса и других језичких извора лексичким информацијама. лематизацију према речнику, и дистрибутивну семантику, односно векторе колокације, ембединге речи, значења и појмова.

Модул се бави случајевима коришћења у лексикографији заснованој на корпусу, корпусној лингвистици и обради природног језика, у комбинацији са основним и другим *OntoLex* модулима.

Помоћна класа *:SrFudKo* дефинисана је да обезбеди лакше руковање и крађу анотацију. Упућивање на фреквенције су коришћене за корпус *срФудКо*, објављен на инстанци "noSketch" (Kilgarriff et al. 2014) коју одржава Друштво за језичке ресурсе и технологије JePTeX.[17]

Уведена је помоћна класа *:SrFudKo_lemma_frek* за фреквенцију свих облика конкретног термина, уз напомену да се термин може састојати од једне или више компоненти. Мотив за увођење наведене класе је компактније кодирање јер се смањује понављање сегмената кода. Поређење фреквенција са референтним општим корпусом српског језика СрпКор2021 (Krstev and Stanković 2023; Škorić and Janković 2024) обезбеђено је кроз евидентирање фреквенција у том корпусу коришћењем објекта *:SrpKor2021*.

```
# korpus fudbala - srpski
:SrFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrFudKo>] .
:SrFudKo_lemma_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
    [a owl:Restriction; owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# opšti korpus savremenog srpskog jezika
:SrpKor2021 a owl:Class;
  rdfs:subClassOf  [a owl:Restriction; owl:onProperty frac:observedIn ;
    owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=SrpKor2021>] .
:SrpKor2021_lemma_freq rdfs:subClassOf
    frac:Frequency, :SrpKor2021, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].

# korpus fudbala - španski
:EsFudKo a owl:Class;
  rdfs:subClassOf  [a owl:Restriction;
    owl:onProperty frac:observedIn ;
  owl:hasValue <https://noske.jerteh.rs/#dashboard?corpname=EsFudKo>] .
:EsFudKo_lemma_freq rdfs:subClassOf
    frac:Frequency, :EsFudKo, [a owl:Restriction;
      owl:onProperty dct:description; owl:hasValue "lemma frequency"].
```

У наставку се илуструју фреквенције лема у корпусу фудбала и општег језика за српски језик. Напоменимо да је један од фактора одређивања термина (енгл. *termhood*) управо претходно поменути појам кључности (енгл. *keyness*) што се рачуна на основу фреквенција у ова два корпуса. Дата су два могућа начина записивања фреквенција.

17. JePTeX

```
# frekvencije na nivou leme  (prvi način)
:le_zxut_86508 frac:frequency
    [a :SrFudKo_token_freq; rdf:value "2313" ] .
:le_zxut_86508 frac:frequency
    [a :SrpKor2021_token_freq; rdf:value "23377" ] .

# frekvencije na nivou leme  (drugi način)
:le_zxut_86508 frac:frequency
   :freq_SrFudKo_le_zxut_86508,
   :freq_SrpKor2021_le_zxut_86508.
:freq_SrFudKo_le_zxut_86508
   a :SrFudKo_token_freq; rdf:value "2313".
:freq_SrpKor2021_le_zxut_86508
   a :SrpKor2021_token_freq;
   rdf:value "23377".

# frekvencije na nivou leme za višečlanu reč
:le_zxuti_karton_216263 frac:frequency
     :freq_SrFudKo_le_zxuti_karton_216263,
     :freq_SrpKor2021_le_zxuti_karton_216263.
```

За шпански језик дајемо примере фреквенција на корпусу *есФудКо*. За CQL (Corpus Query Language) упит [lemma = "amarillo"] добија се 1.517 појављивања, при чему се следеће фреквенције јављају по појединачним облицима: *amarilla* (839), *amarillas* (244), *amarillo* (221), *amarillos* (85), *Amarillas* (55), *Amarilla* (41), *AMARILLAS* (21), *Amarillo* (10) и *AMARILLA* (1). У речнику *ФудЛе* приказују се обједињено фреквенције, без обзира на мала и велика слова, дакле amarillas ће имати 244+55+21=320

```
# frekvencije na nivou leme (prvi način) španski
:le_amarillo_2 frac:frequency
    [a :EsFudKo_token_freq; rdf:value "1517" ] .

# frekvencije na nivou leme  (drugi način)
:le_amarillo_2 frac:frequency :freq_ESFudKo_le_amarillo_2.
:freq_ErFudKo_le_amarillo_2
   a :EsFudKo_token_freq;
   rdf:value "1517".
```

Апсолутне фреквенције добијају се коришћењем CQL израза, док се релативне фреквенције рачунају на милион речи – дељењем са укупним бројем речи у корпусу и множе са милион. Када су у питању фреквенције вишечланих елемената у српском, односно шпанском језику, једно од могућих решења својстава је следеће (Lazarević et al. 2023):

```
:SrFudKo_mwe_freq rdfs:subClassOf frac:Frequency, :SrFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].
```

```
:EsFudKo_mwe_freq rdfs:subClassOf
  frac:Frequency, :EsFudKo,
   [owl:Restriction;
     owl:onProperty dct:description;
     owl:hasValue "mwe frequency"].
```

У наставку су дати примери за фреквенције вишечлане леме „жути картон“. На примеру корпуса *срФудКо* дато је појашњење и упити којима се долази до фреквенција.

Фреквенције лема се добијају упитом “[lemma="žut"][lemma="karton"]”, где осим претходно наведених облика имамо и: „žuti kartoni“ (124), „žuta kartona“ (95), „žutih kartona“, „žutog kartona“ (84), „žute kartone“ (62), „žutim kartonom“ (51), „žuti kartoni“ (15),...

```
# mwe frekvencija lema
:le_zxuti_karton_216263
  frac:frequency [a :SrFudKo_mwe_freq;  rdf:value "1996"];
  frac:frequency [a :SrpKor2021_mwe_freq;  rdf:value "3192"];
  frac:head :le_karton_8815  .
```

На шпанском језику дајемо пример за претрагу по леми CQL упитом [lemma = "tarjeta"][lemma = "amarillo"] који даје 312 погодака: *tarjeta amarilla* (214), *tarjetas amarillas* (75), *Tarjetas amarillas* (12), *Tarjeta amarilla* (10), *tarjeta amarillas* (1). Када дајемо фреквенције рачунамо неосетљивост на мала и велика слова.

```
# mwe lemma frequency za primer na španskom
:fm_tarjeta_amarilla_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "224"];
  frac:head :le_tarjeta_1  .
:le_tarjetas_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "87"];
  frac:head :le_tarjeta_1  .
:le_tarjeta_amarillas_10001
  frac:frequency [a :EsFudKo_mwe_freq;  rdf:value "1"];
  frac:head :le_tarjeta_1  .
```

Треба напоменути да је могуће увођење појединачног објекта за фреквенције, са којим се затим повезују подаци:

```
# frekvencije za španski
:le_tarjeta_amarilla_10001 frac:frequency :freq_le_tarjeta_amarilla_10001.
```

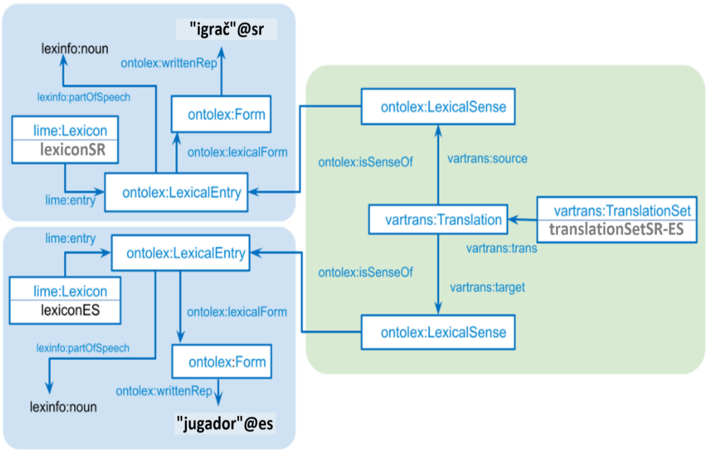
## 4. Преводни еквиваленти у речнику *ФудЛе*

Као што је претходно поменуто, Речник фудбалске терминологије *ФудЛе* састоји се од више различитих компонената, међу којима овде

представљамо: изворни, српски део лексикона, циљни, шпански део и скуп превода (енгл. *TranslationSet*).

Оваква подела се природније уклапа у основну шему Ontolex и модул *vartrans*. Као резултат, два независна једнојезична лексикона на српском и на шпанском језику објављују се као повезани скупови података, заједно са скупом превода који их повезује.

Објављивање речника и за друге језике је могућ по истој шеми. Слика 3 илуструје шему представљања која се користи за превод лексичке јединице „играч“ са српског на шпански у „jugador“, у контексту основног простора имена речника *ontolex* и *vartrans*.




**Слика 3.** Моделирање превода коришћењем *OntoLex* и *vartrans*.

Лексичка јединица (*ontolex:LexicalEntry*) и њена придружена својства се користе за додавање лексичких информација, док их *vartrans:Translation* класа повезује путем лексичког значења *ontolex:LexicalSense*. На пример: својство писана форма (*ontolex:writtenRep*) повезује субјекат класе облик (*ontolex:Form*) и објекат „играч“ на српском језику ili „jugador“ на шпанском.

Постоје и једноставније могућности када се лексичке јединице повезују, без дефинисања посредничког значења. Тада се користи предикат *vartrans:translatableAs*. Међутим, преводни еквиваленти се

успостављају између специфичних значења речи и коришћење класе *vartrans:Translation*, као повезнице између два значења лексичких јединица преко својства *ontolex:LexicalSense*, омогућава да јасно представимо ову чињеницу.

За публиковање речника одабран је следећи образац URI идентификатора: http://{domain}/{type}/{concept}/{reference}, где {type} декларише тип ресурса који се идентификује. На пример: 'id' или 'item' за објекте из реалног света; 'doc' за документе који описују те објекте; 'def' за концепте; 'set' за скупове података; или ниска специфична за контекст, као што је 'authority' или 'dcterms' (Archer, Goedertier, and Loutas 2012). Сличан приступ је коришћен за публиковање Српско-немачког речника *SrpNemLex* (Andonovski 2023).

У случају *ФудЛе* су пројектована три скупа података:

- Скуп података изворног језика (српски): https://llod.jerteh.rs/id/fudle/lexiconSR
- Скуп података циљног језика (шпански): https://llod.jerteh.rs/id/fudle/lexiconES
- Скуп података преводних еквивалената (српско-шпански преводни): https://llod.jerteh.rs/id/fudle/tranSetSR-ES

Следи пример за превод одредница "igrač"@sr i "jugador"@es где су имена објеката прилагођена лакшем разумевању креираних веза између преводних еквивалената на нивоу значења речи.

```
# :lexiconSR a lime:Lexicon .
...
:lexiconSR lime:entry :lexiconSR/igracy-n-sr .
:lexiconSR/igracy-n-sr a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconSR/igracy-n-sr-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconSR/igracy-n-sr-form a ontolex:Form ;
  ontolex:writtenRep "igrač"@sr .
:lexiconES a lime:Lexicon  .
...
:lexiconES lime:entry :lexiconES/jugador-n-es .
:lexiconES/jugador-n-es a ontolex:LexicalEntry ;
  ontolex:lexicalForm :lexiconES/jugador-n-es-form ;
  lexinfo:partOfSpeech lexinfo:noun .
:lexiconES/jugador-n-es-form a ontolex:Form ;
  ontolex:writtenRep "jugador"@es .
...
:tranSetSR-ES a vartrans:TranslationSet   ;
...
:tranSetSR-ES vartrans:trans
  :tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans.
:tranSetSR-ES/igracy_jugador-n-sr-sense a ontolex:LexicalSense ;
```

```
  ontolex:isSenseOf :lexiconSR/igracy-n-sr .
:tranSetSR-ES/jugador_igracy-n-es-sense a ontolex:LexicalSense ;
  ontolex:isSenseOf :lexiconES/jugador-n-es .
:tranSetSR-ES/igracy_jugador-n-sr-sense-jugador_igracy-n-es-sense-trans
  a vartrans:Translation ;
  vartrans:source :tranSetSR-ES/igracy_jugador-n-sr-sense ;
  vartrans:target :tranSetSR-ES/jugador_igracy-n-es-sense .
```

Модул FrAC омогућава да се прикажу и колокације и ембединзи (Chiarcos et al. 2022; Chiarcos et al. 2020), односно различита векторска представљања.

## 5. Закључак

Развој фудбалског речника путем методологије повезаних података показао је значајне предности у односу на традиционалне приступе. Аутоматизација процеса, коришћење корпуса и интеграција различитих дигиталних ресурса омогућили су бржу и ефикаснију израду речника. Дигитални речници нуде многе предности, као што су брзина претраге, стална ажурирања и прилагодљивост корисницима, што их чини незамењивим алатом у савременом добу. Примена техника рачунарске лингвистике и технологија семантичке мреже омогућила је стварање речника који задовољава потребе људи и машина. У будућности очекујемо да се овакви приступи примене на израду речника у другим областима, чиме се додатно унапређује област лексикографије и језичких технологија.

## Захвалница

Истраживање је спроведено уз подршку Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

## Литература

Andonovski, Jelena. 2023. “SrpNemKor and Linked Open Data.” *Infotheca - Journal for Digital Humanities* 23 (1): 33–60. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2023.23.1.2.

Archer, Ph., S. Goedertier, and N. Loutas. 2012. *Study on Persistent URIs, with Identification of Best Practices and Recommendations on the Topic for the MSs and the EC.* Technical report. European Commission.

Bergh, Gunnar, and Sölve Ohlander. 2019. "A hundred years of football English: A dictionary study on the relationship of a special language to general language."

Chiarcos, Christian, Elena-Simona Apostol, Besim Kabashi, and Ciprian-Octavian Truică. 2022. "Modelling Frequency, Attestation, and Corpus-Based Information with OntoLex-FrAC." In *Proceedings of the 29th International Conference on Computational Linguistics,* 4018–4027.

Chiarcos, Christian, Christian Fäth, and Maxim Ionov. 2022. "Unifying morphology resources with OntoLex-morph. a case study in German." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 4842–4850.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Fillmore, Charles J, Miriam RL Petruck, Josef Ruppenhofer, and Abby Wright. 2003. "FrameNet in action: The case of attaching." *International journal of lexicography* 16 (3): 297–332.

Fuertes-Olivera, Pedro A, and Sven Tarp. 2014. *Theory and practice of specialised online dictionaries: Lexicography versus terminography.* Vol. 146. Walter de Gruyter GmbH & Co KG.

Ivanović, Tanja, Ranka Stanković, Branislava Šandrih Todorović, and Cvetana Krstev. 2022. "Corpus-based bilingual terminology extraction in the power engineering domain." *Terminology* 28 (2): 228–263.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. "The Sketch Engine: ten years on." *Lexicography* 1 (1): 7–36.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. "GDEX: Automatically finding good dictionary examples in a corpus." In *Proceedings of the XIII EURALEX international congress,* 1:425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra ...

Kitanović, Olivera, Ranka Stanković, Aleksandra Tomašević, Mihailo Škorić, Ivan Babić, and Ljiljana Kolonja. 2021. "A data driven approach for raw material terminology." *Applied Sciences* 11 (7): 2892.

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

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

Krstev, Cvetana, Ranka Stanković, Ivan Obradović, and Biljana Lazić. 2015. "Terminology Acquisition and Description Using Lexical Resources and Local Grammars." In *Proceedings of the 11th Conference on Terminology and Artificial Intelligence.* Granada, Spain: LexiCon (Universidad de Granada).

Krstev, Cvetana, and Duško Vitas. 2009. "An Effective Methode for Developing a Comprehensive Morphological E-Dicitonary of Compounds." In *Proceedings of the 28th Conference on Lexis and Grammar,* edited by B. Lamiroy, E. Laporte, and T. Kyriakopoulou, 204–212. 29th September - 3rd October 2009, in Arena Romanistica. Bergen: University of Bergen, Department of Foreign Languages.

Lazarević, Jelena, Ranka Stanković, Mihailo Škorić, and Biljana Rujević. 2023. "Football terminology: compilation and transformation into OntoLex-Lemon resource." In *Proceedings of LDK 2023 – 4th Conference on Language, Data and Knowledge.* 12–15 September, Vienna, Austria. Lisabon: NOVA FCSH - CLUNL. https://doi.org/10.34619/srmk-injj.

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.

Mihajlović, Aleksandar. 2003. *Fudbalski rečnik: srpsko-englesko-francusko-španski.* Beograd : A. Mihajlović, 2003 (Beograd : Solidarnost).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26.

Río Gracia, Jorge del, et al. 2023. *Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries (v2.0) Draft Community Group Report.* Technical report. 03 October 2023. October. https://bpmlod.github.io/Bilingual_Dictionaries_Report/.

Robin, Cécile, Atharva Kulkarni, and Paul Buitelaar. 2023. "Identifying FrameNet Lexical Semantic Structures for Knowledge Graph Extraction from Financial Customer Interactions." In *Proceedings of the 12th Global Wordnet Conference,* edited by German Rigau, Francis Bond, and Alexandre Rademaker, 91–100. University of the Basque Country, Donostia - San Sebastian, Basque Country: Global Wordnet Association, January. https://aclanthology.org/2023.gwc-1.11/.

Savary, Agata, Cvetana Krstev, and Duško Vitas. 2007. "Inflectional non compositionality and variation of compounds in French, Polish and Serbian, and their automatic processing." In *Bulag — Bulletin de Linguistique Appliquée et Générale: Les langues slaves et le français : approches formelles dans les études contrastives,* edited by Aleksandra Dziadkiewicz and Izabella Thomas, 73–94. 32. Besançon: Presses Universitaires de Franche Comté.

Schmidt, Thomas. 2009. "The Kicktionary–A multilingual lexical resource of football language." *Multilingual FrameNets in computational lexicography. Methods and applications,* 101–134.

Škorić, Mihailo, and Nikola Janković. 2024. "New Textual Corpora for Serbian Language Modeling." Faculty of Philology, University of Belgrade, *Infotheca* 24 (1).

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries - from File System to lemon Based Lexical Database." In *Proceedings of the 11th LREC – W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018).* Miyazaki, Japan: ELRA.

Stanković, Ranka, Ivan Obradović, Cvetana Krstev, and Duško Vitas. 2011. "Production of morphological dictionaries of multi-word units using a multipurpose tool." In *Proceedings of the Computational Linguistics-Applications Conference,* 77–84.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of the 12th Language Resources and Evaluation Conference,* 3954–3962. Marseille, France: ELRA.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel bidirectionally pretrained taggers as feature generators." *Applied Sciences* 12 (10): 5028.

Vitas, Duško. 1979. "Prikaz jednog sistema za automatsku obradu teksta." *Zbornik radova XIV jugoslovenskog medunarodnog simpozijuma o obradi podataka Informatica* 79:7–101.

Vu, Thuy, Aiti Aw, and Min Zhang. 2008. "Term extraction through unithood and termhood unification." In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II.*

Wu, Jiguang, and Ying Li. 2017. "Research on construction of semantic dictionary in the football field." In *Proceedings of the 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA),* 303–306. IEEE Xplore Full-Text PDF available. IEEE.

Рујевић, Биљана. 2022. "Речници у дигиталном добу – информатичка подршка за српски језик." Докторска дисертација, Универзитет у Београду, Филолошки факултет.

Утвић, Милош. 2011. "Анотација Корпуса савременог српског језика." *ИНФОтека* 12, no. 2 (December): 39–51.

# Израда синтетичког евалуативног скупа података за Српски SentiWordNet

УДК 811.163.41'322.2

Саша Петалинкар
sasa5linkar@gmail.com
ORCID: 0009-0007-9664-3594
*Универзитет у Београду*
*Мултидисциплинарне докторске студије*
*Интелигентни системи*
*Београд, Србија*

**САЖЕТАК:** У раду се представља израда синтетичког скупа за евалуацију Српског SentiWordNet-а која користи велике језичке моделе, посебно модел *Мистрал*. Због недостатка ресурса за анализу сентимента на српском језику, циљ истраживања је премошћавање овог јаза генерисањем скупа за евалуацију и унапређење алата за анализу сентимента на српском. Вредности поларитета сентимента из енглеског SentiWordNet-а аутоматски су мапиране на Српски Ворднет. Како би се ове вредности прецизније прилагодиле српском језику, креиран је посебан скуп за евалуацију. Иницијално је одабрано 500 синсетова из Српског Ворднета, на основу њихове усклађености са senti-pol-sr лексиконом и мапираним вредностима из SentiWordNet-а. Ови синсетови су класификовани према поларитету сентимента коришћењем *Мистрал*-а. Избалансирани подскуп од 75 синсетова насумично је издвојен, додатно профињен финијом градацијом сентимента и ручно прегледан. Резултати показују високу прецизност, приближно 93%.

**КЉУЧНЕ РЕЧИ:** Анализа сентимента, велики језички модели, синтетички скуп података, Српски језик, SentiWordNet

**РАД ПРИМЉЕН:** 23. мај 2024.
**РАД ПРИХВАЋЕН:** 24. јун 2024.

## 1. Увод

Анализа сентимента је процес рачунарског одређивања емоционалног тона иза текста како би се разумели ставови, мишљења и

емоције изражене у њему. Једна од метода за анализу сентимента заснива се на лексиконима сентимента. Лексикони сентимента су специјализовани речници који повезују речи и фразе са вредностима сентимента, омогућавајући аутоматизовану анализу поларитета емоција у тексту (Liu 2010).

Један од проблема који се идентификује код лексикона сентимента је тај што неки изрази могу имати више значења, при чему различита значења речи могу изазивати различите сентименте. Решавање овог проблема подразумева додељивање сентимента према контекстуалном значењу речи, а не према самој речи. Истакнути пример таквог лексикона је SentiWordNet (SWN), који проширује Princeton WordNet (PWN) речник додељивањем оцена сентимента сваком синсету (скуп когнитивних синонима) како би одражавао колективни емоционални тон појма. Овај процес укључује одабир малог броја синсетова са високим поларитетом, након чега следи проширење овог скупа кроз полуаутоматски метод, који идентификује односе унутар ворднета, а који или очувавају или преокрећу поларитет. Добијени проширени скуп се затим користи за обучавање класификатора заснованих на дефиницијама ових синсетова (Baccianella, Esuli, and Sebastiani 2010).

Ворднети за многе језике су међусобно повезани путем међународног језичког индекса (ILI) – синсет у ворднету одређеног језика добија вредност из ILI-а, која представља апстрактан концепт и додељује се синсетовима у другим језицима који лексикализују исти концепт. И EuroWordNet (Vossen 2004) i BalkaNet (Tufis, Cristea, and Stamou 2004) су на овај начин повезани са PWN-ом, а самим тим и са SWN-ом. EuroWordNet укључује ворднете за осам језика: енглески, холандски, шпански, италијански, немачки, француски, чешки и естонски, док BalkaNet проширује ову листу додавањем ворднета за пет додатних језика: бугарски, грчки, румунски, српски и турски (Krstev et al. 2004; Krstev, Koeva, and Vitas 2006; Stanković et al. 2018). Могуће је лако мапирати SWN на било који ворднет који има ILI, што је случај за све ворднете који припадају пројекту Отвореног вишејезичког ворднета (OMW) (Bond and Paik 2012) а који тренутно укључује 200 језика.

Истраживања су показала да се вредности сентимента из SWN-а могу применити на неенглеске језике коришћењем ILI-ја (Denecke 2008). Српски ворднет (СрпВН) садржи вредности сентимента добијене директним мапирањем синсетова коришћењем ILI-ја на SWN (Krstev et al. 2004; Mladenović, Mitrović, and Krstev 2014). СрпВН је коришћен

у креирању хибридног оквира за анализу сентимента на српском језику (Mladenović et al. 2015).

Наша хипотеза је да се овакав лексикон може побољшати заменом мапираних вредности сентимента онима које су репрезентативније за српски језик. Ово је већ спроведено за друге језике, као што су турски (Dehkharghani et al. 2016), арапски (Alhazmi and Black 2013), вијетнамски (Vu and Park 2014), одија (Mohanty, Kannan, and Mamidi 2017), индонежански (Wijayanti and Arisal 2021) и хинди (Bakliwal, Arora, and Varma 2012). Међутим, потребан је евалуациони скуп података – подскуп синсетоба из СрпВН-а већ анотираног са поларитетом сентимента – како би се проценила оваква побољшања за српски језик.

За SWN већ постоји такав евалуациони скуп података: Micro-WNO (Cerini et al. 2007), ручно означен подскуп синсетова из PWN-а, који је јавно доступан.[1] Креирање упоредивог евалуационог скупа података за српски језик ручним путем захтевало би значајан напор, укључујући ангажовање малог броја стручних аннотатора или веће групе мање вештих аннотатора. С обзиром на недостатак таквих ресурса, неопходно је размотрити алтернативни приступ.

Синтетички евалуациони скупови података представљају вештачки креиране колекције података, развијене за тестирање и валидацију рачунарских модела, посебно у областима где стварни подаци могу бити оскудни, пристрасни или превише осетљиви за употребу. Ови скупови података генеришу се путем алгоритама или симулација, са циљем да имитирају статистичке особине стварних података, омогућавајући истраживачима да спроводе робусне евалуације под контролисаним условима (Lu et al. 2024).

Појава ВЕЛИКИХ ЈЕЗИЧКИХ МОДЕЛА (ВЈМ) омогућила је креирање значајно бољих синтетичких скупова података. Показано је да ВЈМ-ови, за потребе задатака из области обраде природног језика (ОПЈ), укључујући анализу сентимента, могу успешно анотирати податке на основу само неколико познатих (виђених) примера (Amatriain 2024; Brown et al. 2020).

У сценарију УЧЕЊА БЕЗ ИЈЕДНОГ ПОКУШАЈА (енгл. Zero-shot learning), модел доноси предикције или анотације без експлицитно виђених примера задатка током тренинга. Ова способност је посебно корисна за анализу сентимента у језицима или контекстима где су анотирани

1. https://github.com/aesuli/Sentiwordnet/blob/master/data/Micro-WNop-WN3.txt

подаци ретки, јер омогућава ВЈМ-у да примени своје већ постојеће знање на нове, невиђене задатке (Amatriain 2024; Brown et al. 2020).

Учење са неколико покушаја (енгл. Few-shot learning), с друге стране, омогућава моделу да из малог броја примера научи како да обавља одређени задатак. Овај приступ је посебно користан за побољшање разумевања модела и повећање његове тачности у специфичним применама, као што је разликовање нијансираних израза сентимента (Brown et al. 2020).

Коришћење упита и одговора (енгл. prompts and reposnse) са ВЈМ-овима омогућава да се ови приступи учења ефикасно примене. Креирањем упита који усмеравају модел ка жељеном излазу, истраживачи могу искористити способности ВЈМ-ова за анализу сентимента. Ово укључује дизајн упита који јасно дефинише задатак, као што је идентификација сентимента датог текста, и омогућавање моделу да генерише одговор на основу свог претходног тренинга и контекста који даје упит. Такав приступ помаже у коришћењу потенцијала ВЈМ-ова за детаљну анализу сентимента, нудећи флексибилан и ефикасан метод за анализу сентимента у различитим скуповима података (Amatriain 2024)

Ово поставља питање да ли би се ВЈМ-ови могли користити не само за креирање скупа за евалуацију података, већ и за анотацију целокупног СрпВН-а са вредностима поларитета сентимента. Одлука да се фокусира на креирање малог скупа за евалуацију проистиче из високих рачунарских трошкова повезаних са анотацијом целокупне мреже.

Примарни циљ овог приступа је побољшање постојећих вредности сентимента у СрпВН-у, које су добијене мапирањем из SWN-а. Фокус је на синсетовима који су у мапирању означени као објективни, иако садрже речи које у лексикону сентимента имају поларитет.

Да би се постигао балансиран узорак погодан за ефикасну евалуацију, посебно у каснијој примени модела машинског учења за класификацију сентимента, насумично је одабрано 500 синсетова. Ови синсетови су обрађени коришћењем модела *Mistral*. Први корак у обради био је категоризација синсетова у три групе: позитивни, негативни и објективни сентимент. Након тога, из сваке категорије је одабран једнак број синсетова за финесу у даљој градацији.

Резултати су прошли кроз процес ручне анотације, где је сваки синсет, заједно са вредностима које је вратио ВЈМ, процењен и

оцењен једноставном ознаком „успех“ или „неуспех“, на основу њихове усклађености са очекиваним вредностима сентимента.

Првобитно је методологија била осмишљена да укључи учење са неколико покушаја за финију градацију сентимента, користећи примере из синсетова који нису изабрани за примарни скуп података — конкретно, оних синсетова који су показивали очигледну усклађеност сентимента у српском и енглеском SWN-у. Међутим, прелиминарни резултати добијени приступом учења без иједног покушаја показали су се задовољавајућим, чиме је компонента учења са неколико покушаја постала непотребна. Због тога се истраживање наставило искључиво са парадигмом учења без иједног покушаја, при чему је *Mistral* модел примењен без претходних специфичних примера за задатак анализе сентимента.

Ручна анотација резултата потврдила је валидност овог поједностављеног приступа. Методологија учења без иједног покушаја показала је изванредну стопу успеха од преко деведесет процената, афирмативно показујући да чак и без укључивања учења са неколико покушаја и додатног контекста који оно пружа, BJM може ефикасно да препозна и класификује сентимент у оквиру одабраних српских синсетова.

Главни BJM коришћен у овом истраживању је *Mistral 7B – Instruct*.[2] То је фино подешена варијанта модела *Mistral 7B*, језичког модела са 7 милијарди параметара, дизајнираног за врхунске перформансе и ефикасност. *Mistral 7B* надмашује модел *Llama 2 13B* на разним бенчмарковима. Нарочито, превазилази *Llama 1 34B* у задацима расуђивања, математике и генерисања кода (Jiang et al. 2023). *Mistral 7B – Instruct* такође надмашује модел Llama 2 13B – Chat како на људским, тако и на аутоматизованим бенчмарковима (Jiang et al. 2023). Објављен под Apache 2.0 лиценцом, овај модел нуди флексибилно средство за истраживаче, са могућношћу да се користи на локалном рачунару без додатних трошкова (Jiang et al. 2023).

Овај рад је организован у неколико кључних поглавља како би се свеобухватно дискутовало о истраживању и резултатима. У поглављу 2 детаљно је описана методологија, укључујући одабир дефиниција синсетова за обраду, специфичне упите коришћене у истраживању, као и примену ових упита. Поред тога, укратко су поменути покушаји коришћења других језичких модела ради поређења. Поглавље 3 приказује резултате, описујући тачност модела са примерима синсетова

2. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

оцењених као исправни и неисправни, као и одговоре који нису припадали задатом скупу дефинисаном инструкцијама упита. На крају, поглавље 4 закључује рад са сажетком налаза и дискусијом о потенцијалним будућим радовима који би могли проширити ово истраживање.

## 2. Методологија

*Senti-pol-sr* је лексикон поларитета за српски језик, анотиран на нивоу речи, а не на нивоу значења речи (Stanković et al. 2022). У нашем истраживању, овај лексикон је коришћен за одабир узорка погодног за анотацију помоћу BJM-а. Ово је постигнуто идентификацијом свих синсетова из СрпВН-а који садрже литерале присутне у лексикону *senti-pol-sr* и истовремено имају неутралну вредност сентимента (0,0), како је мапирано из *SWN*-а.

Детектована су четири главна разлога за неслагања између вредности сентимента у лексикону *senti-pol-sr* и оних изведених из *SWN*-а. Прво, док реч може преносити поларизујући сентимент, стварно значење у којем се користи можда не носи тај сентимент. На пример, синсет `ENG30-06828389-n` (деф. „карактер налик на звезду (*) који се користи у штампаним текстовима“) у Српском Ворднету садржи литерал „звезда“, која у другим синсетовима може имати позитивне конотације. Друго, вредности сентимента у *SWN*-у, генерисане методама машинског учења, могу бити нетачне. На пример, синсет `ENG30-02585489-v` (деф. „изазвати патњу“) који је очигледно веома негативан. Треће, ако синсет нема одговарајући синсет у *PWN*-у, као што је `BILI-00000941` (деф, „Гласно изражавати жалост, запевати, тужити.“), уобичајено му се додељује поларитет (0,0).

Најинтересантнија могућност је да се концепт сматра објективним на енглеском, док на српском језику носи сентимент. Такви синсетови још нису пронађени.

Почетна анализа идентификовала је 2.956 синсетова од укупно 25.320 унутар СрпВН-а(Mladenović, Stanković, and Krstev 2017) који су садржали литерале анотиране са јасним поларитетом у лексикону *senti-pol-sr* и имали вредност поларитета (0,0), од чега 1.511 показује позитиван сентимент, а 1.445 негативан. С обзиром на велики обим, обрада свих ових синсетова помоћу BJM-а није била изводљива, па је насумично одабран узорак од 500 синсетова за даљу анализу.

Ради креирања уравнотеженог скупа узорака, дефиниције из овог насумичног узорка су обрађене коришћењем *LangChain Python* библиотеке, моћног алата за креирање, експериментисање и анализирање језичких модела и агената (Chase 2022). *LangChain Python* библиотека функционише као интерфејс за модуле више великих језичких модела. Овај пројекат је користио омотаче пројектоване за Hugging Face Hub и Transformer библиотеке. Процедура коришћена у овом истраживању садржала је само један шаблон упита са једном улазном променљивом – дефиницијом синсета, и *Mistral 7B – Instruct* модел преузет са *Hugging Face Hub*-а. Модел је извршен на локалном рачунару. Библиотека *LangChain* омогућава да се упитни шаблони са променљивама, које су означене витичастим заградама, попуне вредностима тих променљивих и проследе као упит BJM модулу, који затим враћа одговор.

Користећи одговарајући упит као што је приказано на слици 1, узорак је подељен на оне означене као позитивне, негативне, објективне и оне који нису исправно означени. Било је 290 објективних, 102 негативна, 33 позитивна и 75 погрешно означених синсетова.

Да би се искористиле пуне могућности BJM-а за анализу сентимента, упит је пажљиво пројектован да обухвати три кључна елемента: доделу улога (енгл. *role-play*), јасне инструкције и очекиване исходе.

1. Додела улога: инструкција BJM-у као експерту: Упит започиње сценаријом играња улога, где се BJM-у даје инструкција да преузме улогу експерта за анализу сентимента (слика 1). Овај приступ је усвојен како би се модел усмерио ка аналитичком и фокусираном испитивању текста, ослањајући се на своје опсежно унапред обучено знање и разумевање нијанси анализе сентимента.
2. Јасне инструкције: суштина ефикасне комуникације са BJM-ом лежи у јасноћи инструкција. Упит експлицитно описује задатак, водећи BJM да идентификује и анализира сентимент изражен у датом делу текста (дефиницији синсета). Ова јасноћа осигурава да су аналитичке способности модела усмерене ка тачном процењивању сентимента, минимизујући двосмисленост у његовим одговорима.
3. Очекивани исходи: да би се додатно усавршио излаз модела, упит прецизира очекиване исходе анализе. Наводи жељени формат одговора, било да је то класификација сентимента (позитиван, негативан, неутралан) или нијансиранија оцена сентимента.

Упит коришћен за класификацију је усавршен путем експерименталног тестирања на мањем подскупу дефиниција синсета. Упоредне анализе инструкција упита на енглеском и српском језику откриле су да су упити на српском језику давали тачније резултате. Како би се обезбедило свеобухватно покривање потенцијалних одговора, број токена у излазу је постављен на пет.

Даље испитивање показало је постојање дупликата унутар скупа, због тога што неки синсетови садрже више литерала који су такође речи из senti-pol-sr лексикона, као што је синсет `ENG30-01375831-a` „славан, познат, чувен, значајан, реномиран, признат, прослављен: опште прихваћен и уважен“. Важно је напоменути да се сви осим два литерала („реномиран“, „признат“) из овог синсета појављују у лексикону, и означени су као позитивни. Након уклањања дупликата, остало је 279 објективних, 97 негативних и 27 позитивних синсетова. У овом кораку није било потребе за бројањем неправилно означених синсетова, јер уклањање дупликата из њих нема практичан значај.

За детаљнију анализу сентимента, насумично је одабран подскуп од по 25 синсетова из сваке категорије сентимента, формирајући такозвани „балансирани узорак“.

```
Kao ekspert za analizu sentimenta, analizirajte
sledeći tekst na srpskom jeziku i odredite njegov
sentiment.
   Sentiment treba da bude striktno klasifikovan
kao "pozitivan", "negativan", ili "objektivan".
Nijedan drugi odgovor neće biti prihvaćen.
   Tekst: {text}
   Sentiment:
```

**Слика 1.** Шаблон упита за одређивање поларитета.

Балансирани узорак је даље обрађен кроз два шаблона упита за нијансирану обраду сентимента, један за позитиван и један за негативан сентимент (слике 2 и 3). Ови шаблони упита су пројектовани експериментално, кроз итеративно тестирање разних опција и постепено прилагођавање текста, користећи мале скупове дефиниција синсетова из

СрпВН-а. На пример, понављање реченице „Ниједан други одговор неће бити прихваћен“ два пута значајно је смањило број одговора ван оквира.

Јединствени упит за одређивање обе вредности није одабран како би се очувала доследност са структуром SWN-а, где синсет може имати позитивне (POS) и негативне (NEG) вредности изнад нуле, све док њихов збир није већи од један. Да би се прилагодила очекивана дужина излазних стрингова, број излазних токена је повећан на девет.

Током ове фазе, тестирани су и други BJM модели на малом узорку – *GPT2-ORAO* (Škorić 2024), *Llama-2-7b* (Touvron et al. 2023), *alpaca-serbian-7b-base* и *WizardLM-1.0-Uncensored-Llama2-13b*. Идеја је била да се упореде различити резултати, симулирајући анотацију од стране више анотатора. Међутим, резултати су довели до њиховог искључења из даљег рада због превеликог броја неправилно означених синсетова и недостатка финије градације.

```
Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima pozitivan sentiment.
Sentiment treba da bude striktno klasifikovan kao "nije pozitivan", "slabo pozitivan", "umereno pozitivan", "veoma pozitivan", ili "ekstremno pozitivan". Nijedan drugi odgovor neće biti prihvaćen.
Nijedan drugi odgovor neće biti prihvaćen.
Tekst: {text}
Pozitivan sentiment:
```

**Слика 2.** Шаблон упита за фину ознаку позитивног сентимента.

За сваку дефиницију у балансираном узорку, додељене су две вредности на тај начин. Интенција шаблона упита и дизајн ограничавају одговоре на следећи сет одговора:

– За позитиван сентимент:
  - „није позитиван“
  - „слабо позитиван“
  - „умерено позитиван“

*Kao ekspert za analizu sentimenta, analizirajte sledeći tekst na srpskom jeziku i odredite da li ima negativan sentiment.*
*Sentiment treba da bude striktno klasifikovan kao "nije negativan", "slabo negativan", "umereno negativan", "veoma negativan", ili "ekstremno negativan". Nijedan drugi odgovor neće biti prihvaćen.*
*Nijedan drugi odgovor neće biti prihvaćen.*
*Tekst: {text}*
*Negativan sentiment:*

**Слика 3.** Шаблон упита за фину ознаку негативног сентимента.

- „веома позитиван“
- „екстремно позитиван“

– За негативан сентимент:
- „није негативан“
- „слабо негативан“
- „умерено негативан“
- „веома негативан“
- „екстремно негативан“

Резултирајући скуп података је сачуван у форми датотеке са вредностима одвојеним зарезима (CSV).[3] Колоне у тој датотеци су следеће:

**ILI:** Међуезички Индекс, који повезује синсет са његовим еквивалентима у WordNet-има других језика.
**Definition:** Глоса синсета, која пружа његово значење или објашњење.
**Lemma_names:** Литерали садржаних у синсету.
**Sentiment_SWN:** Вредност сентимента мапирана из SWN-а, која показује оригинални скор сентимента у енглеској верзији.

3. https://github.com/sasa5linkar/SWN-synth-eval-set/blob/main/balanced_sample2.csv

**Sentiment_lexicon** : Вредност сентимента изведена из лексикона сентимента senti-pol-sr креираног за српски језик, одражавајући локалне нијансе сентимента.

**Sentiment_sa** : Иницијална класификација од стране великог језичког модела, категоризујући сентимент као позитиван, негативан или неутралан.

**Sentiment_sa_positive** : Фино класификовање сентимента за позитиван сентимент, означавајући степен позитивности од „није позитиван“ до „екстремно позитиван“.

**Sentiment_sa_negative** : Фино класификовање сентимента за негативан сентимент, означавајући степен негативности од „није негативан“ до „екстремно негативан“.

С обзиром на мали узорак од 75, аутор је могао спровести ручну евалуацију целог скупа података. Овај процес је унапређен коришћењем *Data Wrangler*[4] екстензије у *Visual Studio Code*-у, која је алатка за визуализацију и чишћење података и добро се интегрише са *VS Code*-ом и *VS Code Jupyter Notebooks*-ом. Алатка пружа интерактивни интерфејс за преглед и анализу података, нуди информативне статистике и визуелне приказе, и може убрзати чишћење података аутоматским генерисањем Pandas скрипти за модификацију података. Користила се за пажљиво испитивање дефиниција синсетова и њихово поређење са детаљним повратним информацијама о сентименту.

## 3. Резултати

| **Излаз** | **Број у sentiment_sa_negative** |
|---|---|
| екстремно негативан | 1 |
| није негативан | 60 |
| умерено негативан | 1 |
| умјерено негативан | 2 |
| веома негативан | 11 |
| **Укупно** | 75 |

**Табела 1.** Негативан сентимент.

4. microsoft/vscode-data-wrangler (github.com)

| **Излаз** | **Број у sentiment_sa_positive** |
|---|---|
| Није Ова из | 1 |
| Није позитиван | 39 |
| Nijedan O | 6 |
| Nijedan Текст | 1 |
| Учимо се држати | 1 |
| Умерено позитиван | 3 |
| Умерено позитиван The | 2 |
| Умјерено позитиван | 6 |
| веома позитиван | 12 |
| Веома позитиван O | 1 |
| Веома позитиван R | 2 |
| Веома позитиван Ov | 1 |
| **Укупно** | 75 |

**Табела 2.** Позитиван сентимент

Резиме излаза генерисаних од стране BJM-а коришћењем шаблона упита за фину градацију сентимента приказан је у табелама 1 и 3.. Подаци у табелама показују да излаз BJM-а није био ограничен како је назначено у шаблону упита, али је остао унутар граница које се лако могу исправити. Излаз генерисан од негативног шаблона упита у великој мери се поклапао са предложеним излазом, уз само мање дијалекатске варијације. Насупрот томе, излаз из позитивног шаблона упита укључивао је један бесмислен одговор, „учимо се држати“, као и неке одговоре који се могу лабаво тумачити као не-позитивни. Током ручне провере, сви ови одговори су класификовани као не-позитивни.

Додатно, мање типографске грешке, као што су додатна слова, погрешни падежи или грешке у великим словима, занемарене су током ручне провере. Само једна особа је спровела евалуацију. Надаље, у излазима нису пронађени примери благо позитивних или благо негативних синсетова.

За садржај су два синсета очигледно неправилно означена, а три су била сумњива.

Очигледно неправилно су означени синсетови:

– `BILI-00000941`, ‘нарицати: Гласно изражавати жалост, запевати, тужити.’ означен је као потпуно објективан (нити позитиван нити

негативан), док очигледно носи негативан сентимент, штавише, изразито негативан.
– `ENG30-04525038-n`, ‘баршун, сомот, плиш, кадифа: Свиленкаста, густа, чупава тканина, равна са полеђина.’, који је означен као благо позитиван. Као врста текстила, требало би да буде објективан.

Синсетови који се сматрају сумњивим су:

– `ENG30-01215137-v`, ‘ухапсити, затворити, скембати: Ставити у притвор, као осумњичене криминалце; о полицији.’ означен као благо негативан, док га је евалуатор оценио као изразито негативан.
– `ENG30-00309647-n`, ‘експедиција: Путовање организовано са специјалним циљем.’ означен као веома позитиван, док га је евалуатор сматрао неутралним.
– `ENG30-03135152-n`, ‘крст: крст као знамење хришћанства’ означен као веома позитиван, док га је евалуатор сматрао неутралним.

Од већине синсетова који су тачно означени током нашег процеса анализе сентимента, овде представљамо избор илустративних примера. Ови примери би такође требало да демонстрирају критеријуме коришћене у евалуацији приликом разматрања исправне ознаке у класификацијама на позитивне, негативне и неутралне сентименте. Неки примери су:

– `ENG30-01220336-n`: Синсет повезан са акцијама попут „клевета“, „клеветање“ и „омаловажавање“, описан као „оштар напад на чију личност или добро име“. Сентимент овог синсета оцењен је као „Није позитиван“, што одражава одсуство позитивног сентимента, и прецизније, „Екстремно негативан“, тачно хватајући негативне конотације описаних акција.
– `ENG30-06828389-n`: Овај синсет се односи на „карактер налик на звезду (*) који се користи у штампаним текстовима“, са литералима „астериск“, „звездица“ и „звезда“. Сентимент за овај синсет је прецизно класификован као „Није позитиван“ и „Није негативан“, указујући на његову објективну природу без инхерентног позитивног или негативног сентимента.
– `ENG30-10407310-n`: Односи се на „онај који воли и брине о својој земљи“, са литералима „домољуб“, „патриота“ и „родољуб“. Сентимент овог синсета је фино оцењен као „Веома позитиван“ и „Није негативан“, ефикасно хватајући позитивне конотације повезане са патриотизмом.

## 4. Закључак

Анализа излаза генерисаног од стране BJM-а, конкретно *Mistral* модела, показује да је 70 од 75 одговора испунило критеријуме прихватљивости дефинисане за ово истраживање. Овај резултат, који представља значајну већину скупа података, наглашава поузданост и ефикасност *Mistral* модела у извођењу класификације сентимента за српски језик. Са приближном стопом успеха од 93,3%, може се са сигурношћу закључити да је скуп података и употребљив и довољног квалитета за намеравану сврху евалуације предложених корекција поларитета сентимента унутар Српског WordNet-а.

Неуспех других тестираних модела може бити резултат упита оптимизованих за *Mistral* модел. Креирање упита за сваки модел појединачно могло би омогућити употребу више модела за постизање бољег квалитета.

Значајно подручје за будућа истраживања укључује проширење шаблона упита са специфичним примерима, прелазећи из тренутног приступа учења без иједног покушаја на парадигму учења са неколико покушаја. Ова модификација се очекује да побољша капацитет модела за нијансирану градацију сентимента, нудећи прецизнију и контекстуално свеснију анализу. Посебно се препоручује да се ово унапређење селективно примени на шаблоне за фину градацију сентимента, где је потенцијал за повећану тачност и осетљивост на језичке суптилности најизраженији (Brown et al. 2020).

Даље, истраживање напредних методологија за креирање упита, као што је приступ „ланац мисли“ (енгл. chain-of-thought), заслужује пажњу. Ова техника, олакшавајући структуриран и логичан ток мисли у оквиру обраде модела, може значајно побољшати способност модела да извади релевантне податке из одговора. Потенцијал таквих напредних стратегија за превазилажење инхерентних изазова у тачном идентификовању и класификацији израза сентимента у сложеним језичким контекстима представља обећавајући правац за истраживање (Amatriain 2024).

Ово истраживање представља иницијално испитивање могућности коришћења модела *Mistral* за генерисање додатних синтетичких скупова података за српски језик, посебно у областима где су ресурси оскудни или их је тешко прибавити из природних корпуса. Успешна примена модела *Mistral* за анализу сентимента наглашава његов потенцијал као вредног алата у превазилажењу ових изазова.

## Литература

Alhazmi, Samah, and William Black. 2013. "Arabic SentiWordNet in Relation to SentiWordNet 3.0." *John McNaught International Journal of Computational Linguistics (IJCL),* no. 4, 1.

Amatriain, Xavier. 2024. *Prompt Design and Engineering: Introduction and Advanced Methods.* ArXiv:2401.14423 [cs], January. Accessed February 7, 2024. https://doi.org/10.48550/arXiv.2401.14423. http://arxiv.org/abs/2401.14423.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "SENTIWORDNET 3.0: An Enhanced Lexical Resource." In *7th LREC,* 2200–2204. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf.

Bakliwal, Akshat, Piyush Arora, and Vasudeva Varma. 2012. "Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification." In *International Conference on Language Resources and Evaluation.* https://api.semanticscholar.org/CorpusID:1657748.

Bond, Francis, and Kyonghee Paik. 2012. "A Survey of Wordnets and Their Licenses." In *Proceedings of the 6th Global WordNet Conference (GWC 2012),* 64–71. Matsue, Japan.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. *Language Models are Few-Shot Learners.* ArXiv:2005.14165 [cs], July. Accessed February 8, 2024. https://doi.org/10.48550/arXiv.2005.14165. http://arxiv.org/abs/2005.14165.

Cerini, S., V. Compagnoni, A. Demontis, Maicol Formentelli, and C. Gandini. 2007. "Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining." *Language Resources and Linguistic Theory* (January): 200–210.

Chase, Harrison. 2022. *LangChain,* October. https://github.com/langchain-ai/langchain.

Dehkharghani, Rahim, Y. Saygin, B. Yanikoglu, and Kemal Oflazer. 2016. "SentiTurkNet: a Turkish polarity lexicon for sentiment analysis." *Language Resources and Evaluation* 50:667–685. https://doi.org/10.1007/s10579-015-9307-6.

Denecke, Kerstin. 2008. "Using SentiWordNet for multilingual sentiment analysis." In *International Conference on Data Engineering,* 507–512. https://doi.org/https://doi.org/10.1109/ICDEW.2008.4498370.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, et al. 2023. *Mistral 7B.* ArXiv:2310.06825 [cs], October. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2310.06825. http://arxiv.org/abs/2310.06825.

Krstev, Cvetana, Svetla Koeva, and Duško Vitas. 2006. "Towards the Global Wordnet." In *Conference Abstracts of the First International Conference of Digital Humanities Organisations (ADHO) Digital Humanities 2006,* 114–117. Paris-Sorbonne, July. http://poincare.matf.bg.ac.rs/~cvetana/biblio/DigHum06-KKV.pdf.

Krstev, Cvetana, Gordana Pavlović-Lažetić, Duško Vitas, and Ivan Obradović. 2004. "Using Textual and Lexical Resources in Developing Serbian Wordnet." Publisher: Romanian Academy, Publishing House of the Romanian Academy, *Romanian Journal of Information Science and Technology* 7 (1-2): 147–161. http://poincare.matf.bg.ac.rs/~cvetana/biblio/RJIST-MATF.pdf.

Liu, Bing. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing, Second Edition,* edited by Fred J. Damerau and Nitin Indurkhya, 629–661. Chapman & Hall/CRC.

Lu, Yingzhou, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2024. *Machine Learning for Synthetic Data Generation: A Review.* ArXiv:2302.04062 [cs] version: 6, June. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2302.04062. http://arxiv.org/abs/2302.04062.

Mladenović, Miljana, Jelena Mitrović, and Cvetana Krstev. 2014. "Developing and Maintaining a WordNet: Procedures and Tools." In *Proceedings of the Seventh Global WordNet Conference 2014,* edited by Heili Orav, Christiane Fellbaum, and Piek Vossen, 55–62. Tartu, Estonia: University of Tartu, January. ISBN: 978-9949-32-492-7.

Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2015. "Hybrid sentiment analysis framework for a morphologically rich language." *Journal of Intelligent Information Systems* 46 (3): 599–620. Accessed January 1, 2023. https://doi.org/10.1007/s10844-015-0372-5.

Mladenović, Miljana, Ranka Stanković, and Cvetana Krstev. 2017. "A WordNet Ontology in Improving Searches of Digital Dialect Dictionary." In *New Trends in Databases and Information Systems: ADBIS 2017 Short Papers and Workshops - SW4CH (Semantic Web for Cultural Heritage),* 767:373–383. Lecture Notes in Computer Science. Nicosia, Cyprus: Springer International Publishing, September.

Mohanty, Gaurav, Abishek Kannan, and Radhika Mamidi. 2017. "Building a SentiWordNet for Odia." In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* edited by Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot, 143–148. Copenhagen, Denmark: Association for Computational Linguistics, September. https://doi.org/10.18653/v1/W17-5219. https://aclanthology.org/W17-5219/.

Škorić, Mihailo. 2024. "New Language Models for Serbian." Publisher: Zajednica biblioteka univerziteta u Srbiji, Beograd, *Infotheca - Journal for Digital Humanities* 24 (1). https://doi.org/https://doi.org/10.48550/arXiv.2402.14379. https://arxiv.org/abs/2402.14379.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *2nd Workshop on Sentiment Analysis and Linguistic Linked Data (SALLD),* 31–38. Marseille: ELRA, June. Accessed August 21, 2023. https://aclanthology.org/2022.salld-1.6.

Stanković, Ranka, Miljana Mladenović, Ivan Obradović, Marko Vitas, and Cvetana Krstev. 2018. "Resource-based WordNet Augmentation and Enrichment." In *Proceedings of the Third International Conference on Computational Linguistics in Bulgaria (CLIB 2018),* 104–114. Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences, May. https://aclanthology.org/2018.clib-1.14/.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models.* ArXiv:2307.09288 [cs], July. Accessed August 1, 2024. https://doi.org/10.48550/arXiv.2307.09288. http://arxiv.org/abs/2307.09288.

Tufis, D., D. Cristea, and S. Stamou. 2004. "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview." Publisher: Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania, *Romanian Journal of Information Science and Technology* 7 (1-2): 9–43.

Vossen, Piek. 2004. "EuroWordNet: A Multilingual Database of Autonomous and Language-Specific WordNets Connected via an Inter-Lingual Index." _Eprint: https://academic.oup.com/ijl/article-pdf/17/2/161/2912694/ech041.pdf, *International Journal of Lexicography* 17, no. 2 (June): 161–173. ISSN: 0950-3846. https://doi.org/10.1093/ijl/17.2.161. https://doi.org/10.1093/ijl/17.2.161.

Vu, Xuan-Son, and Seong-Bae Park. 2014. "Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary." *ArXiv* abs/1412.8010.

Wijayanti, Rini, and Andria Arisal. 2021. "Automatic Indonesian Sentiment Lexicon Curation with Sentiment Valence Tuning for Social Media Sentiment Analysis." *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 20:1–16. https://doi.org/10.1145/3425632.

# Нови текстуални корпуси за моделовање српског језика

УДК 811.163.4'322.2

**САЖЕТАК:** Овај рад ће представити текстуалне корпусе за српски (и српскохрватски) који се могу користити за тренирање великих језичких модела, а који су јавно доступни на једном од неколико значајних веб репозиторијума. Сваки корпус ће бити класификован помоћу више метода и његове карактеристике ће бити детаљно описане. Поред тога, рад ће представити три нова корпуса: нови кровни веб-корпус за српскохрватски, нови висококвалитетни корпус заснован на докторским дисертацијама похрањеним у Националном репозиторијуму докторских дисертација са свих универзитета у Србији, и паралелни корпус превода сажетака из истог извора. Јединственост старих и нових корпуса биће оцењена путем стилометријских метода заснованих на фреквенцији, и укратко ће се дискутовати о резултатима.

**КЉУЧНЕ РЕЧИ:** корпуси, српски језик, језички модели, евалуација.

**РАД ПРИМЉЕН:** 12. април 2024.
**РАД ПРИХВАЋЕН:** 15. мај 2024.

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
ORCID: 0000-0003-4811-8692
*Универзитет у Београду*
*Рударско-геолошки факултет*
*Београд, Србија*

Никола Јанковић
nikolajankovickv@gmail.com
ORCID: 0000-0003-3484-4220
*Универзитет у Београду*
*Филолошки факултет*
*Београд, Србија*

## 1. Увод

С наглим повећањем доступних текстуалних података кроз феномен *Big Data* почетком двадесет првог века, убрзо се увидело да се ти подаци могу користити за изградњу корпуса за моделовање природног језика. Подаци са интернета, чија количина брзо расте, најпре су коришћени као додатак подацима из књига, који су споријe расли, али, са повећаним интересовањем за количину, полако али сигурно достигли су и надмашили удео података заснованим на књигама у разним корпусима

за тренирање језичких модела. Данас, већина јавно доступних корпуса користи податке са интернета, углавном због лабавијих ограничења ауторских права.

У контексту овог истраживања, категорисаћемо скупове података у следеће категорије на основу њиховог порекла:

$O_1$ Веб-корпуси – корпуси који садрже текстове прикупљене са отвореног интернета, првенствено коришћењем аутоматског сакупљања HTML страница;

$O_2$ Корпуси високог квалитета – корпуси који садрже текстове из уџбеника или сличних извора, односно научне публикације свих врста (научне монографије и чланци у научним часописима, радови са конференција/зборника, тезе итд.) и других објављених нелитерарних дела (владини и правни текстови који нису присутни на интернету, филозофски текстови, кулинарски рецепти итд.) прикупљени првенствено из дигитално насталих докумената или сканираних физичких копија;

$O_3$ Литерарни корпуси – корпуси који садрже објављене литерарна дела укључујући митологију и религиозне текстове;

$O_4$ Синтетички корпуси – корпуси који садрже текстове који нису са Интернета, а који су креирани машинским путем коришћењем метода генерисања природног језика или машинског превођења;

$O_5$ Мешовити корпуси – корпуси креирани коришћењем текстова који потичу из мешавине наведених и других извора.

Осим очигледне разлике у изворној грађи, ове категорије се разликују и по процесу креирања текстова које садрже: док су текстови из прве три групе ($O_1$-$O_3$) углавном креирани од стране људи (нико не може гарантовати да преузета HTML страница није машински генерисан текст), само друга и трећа група садрже оне текстове који су нужно уређени, тј. прочитани и исправљени пре објављивања, што захтева време, али осигурава виши квалитет. Текстови креирани од стране машина ($O_4$) обично се припремају најбрже од свих, али су повезани са нижим квалитетом.

Још једна важна класификација је облик корпуса. С тим у вези, корпусе класификујемо на следећи начин:

$F_1$ Корпуси необележеног текста – корпуси који садрже само текстове који се могу користити за претренирање великих језичких модела као што су *BERT*[1] (Devlin et al. 2018) и *GPT*[2] (Radford et al. 2018);

$F_2$ Обележени корпуси – корпуси где су токени, реченице или други сегменти обележени, нпр. текстови обележени према врсти речи, или реченице обележене на основу сентимента– ови корпуси се обично користе за фино подешавање модела за задатке обележавања текста;

$F_3$ Паралелни корпуси – корпуси где свака реченица (или други сегмент) има специфичан пар, нпр. превод на други језик или одговор на питање – ови корпуси се обично користе за тренирање модела са генеративном компонентом као што је *T5*[3] модел (Raffel et al. 2020).

У наредном одељку, рад ће се фокусирати на корпусе српског и српскохрватског језика прибављене из три значајна веб извора за корпусе и скупове података: *Hugging Face*,[4] *CLARIN virtual language repository*,[5] и *European Language Grid*.[6] Други појединачни репозиторијуми (нпр. *GitHub* пројекти) нису претраживани. Основне информације о корпусима преузетим са ових платформи биће представљене у одељку 2., новоприпремљени корпуси биће представљени у одељку 3., а експеримент везан за процену њихове јединствености биће представљен у одељку 4. Коначно, дискусија о резултатима ове процене налази се у одељку 5.

## 2. Доступни корпуси

Као што је већ поменуто, у овом одељку биће представљена листа корпуса која је састављена прегледом три веб извора: *Hugging Face* (HF), *CLARIN virtual language repository* (VLO) и *European Language Grid* (ELG), с тим што су језици корпуса ограничени на домен српско-хрватског макројезика, односно српски, црногорски, хрватски

1. Bidirectional Encoder Representations from Transformers
2. Generative Pre-trained Transformer
3. Text-To-Text Transfer Transformer
4. huggingface.co, највеће чвориште на Интернету за објављивање великих језичких модела.
5. www.clarin.eu дигитална инфраструктура која пружа податке, алате и сервисе за подршку истраживањима заснованим на језичким ресурсима.
6. live.european-language-grid.eu, платформа за све европске језичке технологије настале из MetaNet-a.

и босански, јер тренирање са сродним језицима може бити корисно у одређеним сценаријима, посебно за задатке који укључују вишејезично разумевање или превођење. Преузети корпуси су категорисани и по форми (примарна класификација) и по пореклу (секундарна класификација). За корпусе необележеног текста ($F_1$), минимални услов за укључивање био је тридесет милиона речи за веб-корпусе ($O_1$), и три милиона речи за остале ($O_2$-$O_5$). Пошто је већина корпуса у овој категорији веб порекла, спроведено је и кратко истраживање о дедупликацији. Када су у питању обележени ($F_2$) и паралелни корпуси ($F_3$), критеријум за величину је био постављен на 3000 реченица, јер су ови ресурси много ређи и обично се користе само за фино подешавање.

### 2.1 Јавно доступни корпуси необележеног текста

Табеле 1 и 2 детаљно описују двадесет четири преузета корпуса необележеног текста. Прва табела се фокусира само на српске корпусе, док друга представља остале корпусе у оквиру српско-хрватског макројезика, укључујући неколико кровних корпуса (произведених обједињавањем и дедупликацијом неколико других корпуса).

| **Назив** | **Јез.** | **Порекло** | **Вел.** | **Издавач** | **Чвор.** |
|---|---|---|---|---|---|
| srWaC | sr | web | 493 | ReLDI | VLO |
| cc100_sr | sr | web | 711 | Conneau et al. | HF |
| mC4-sr | sr | web | 800 | Google | HF |
| OSCAR-sr | sr | web | 632 | OSCAR proj. | HF |
| CLASSLA-sr | sr | web | 752 | CLASSLA | VLO |
| MaCoCu-sr | sr | web | 2.491 | Bañón et al | VLO |
| PDRS1.0 | sr | web | 602 | ISJ | VLO |
| SrpKorNews | sr | web | 468 | JeRTeh | HF |
| SrpELTeC | sr | literary | 5,3 | JeRTeh | HF |

**Табела 1.** Корпуси необележеног текста искључиво на српском језику, доступни на истраженим чвориштима скупова података. Величина је представљена у милионима речи (M)

| Назив | Јез. | Порекло | Вел. | Издавач | Чвор. |
|---|---|---|---|---|---|
| meWaC | cnr | web | 80 | ReLDI | VLO |
| hrWaC | hr | web | 1.250 | ReLDI | VLO |
| bsWaC | bs | web | 256 | ReLDI | VLO |
| cc100_hr | hr | web | 2.880 | Conneau et al. | HF |
| CLASSLA-hr | hr | web | 1.341 | CLASSLA | VLO |
| CLASSLA-bs | bs | web | 534 | CLASSLA | VLO |
| hr_news | hr | web | 1.433 | CLASSLA | HF |
| MaCoCu-cnr | cnr | web | 161 | Bañón et al | VLO |
| MaCoCu-hr | hr | web | 2.363 | Bañón et al | VLO |
| MaCoCu-bs | bs | web | 730 | Bañón et al | VLO |
| riznica | hr | mixed | 87 | IHJJ | VLO |
| HPLT 1.2-sh | mixed | web | 10.030 | HPLT proj. | HF |
| BERTić-data | mixed | ∼web | 8.388 | CLASSLA | VLO |
| MaCoCu-hbs | mixed | web | 5.490 | CLASSLA | HF |
| XLM-R-BERTić-data | mixed | ∼web | 11.539 | CLASSLA | HF |

**Табела 2.** Српско-хрватски корпуси необележеног текста доступни на истраженим чвориштима скупова података, који нису искључиво на српском језику. Величина је представљена у милионима речи (M).

Осим четири корпуса, остали корпуси садрже искључиво материјал преузет са Интернета. Један корпус је литерарни, један је мешовити, а преостала два су агрегирана и класификована као готово веб-корпуси (∼web), пошто укључују поменути мешовити корпус, *riznica* (Ćavar and Brozović Rončević 2012).

Већина ових корпуса је производ неколико засебних подухвата. Корпуси *srWac*, *meWac*, *hrWac* и *bsWac* настали су преузимањем садржаја веб страница са домена највишег нивоа за поменута четири језика (Ljubešić and Klubička 2014; Ljubešić and Erjavec 2021), што је резултирало са више од две милијарде речи. Процес је касније поновљен како би се произвели *CLASSLA-sr*, *CLASSLA-hr* и *CLASSLA-bs*, прикупивши преко 2,5 милијарди речи (Ljubešić and Lauc 2021).

Додатних шест корпуса је изведено из скупа података *Common Crawl*: *OSCAR-sr* (632M) је изведен из скупа података пројекта *OSCAR* (Suárez, Sagot, and Romary 2019), *mC4-sr* (800M) је изведен из

вишејезичног скупа података C4, који је креирала компанија Гугл (Xue et al. 2021), *HPLT 1.2-sh* (преко 10 милијарди речи) је добијен из скупа података пројекта HPLT (Aulamo et al. 2023), док су *cc100_sr* и *cc100_hr* (преко 3,5 милијарди речи) изведени из cc100 скупа података (Conneau et al. 2019). Треба напоменути да су неки од ових подухвата произвели додатне корпусе који нису укључени у ову студију јер нису испунили минимални услов од три милиона речи.

MaCoCu пројекат сакупљања веб страница (Banón et al. 2022; Kuzman and Ljubešić 2023) произвео је *MaCoCu-sr* (2,5 милијарде речи), *MaCoCu-cnr* (151M), *MaCoCu-hr* (2,2 милијарде), *MaCoCu-bs* (686M), и њихов дедупликовани агрегат, *MaCoCu-hbs* (5,5 милијарди речи).

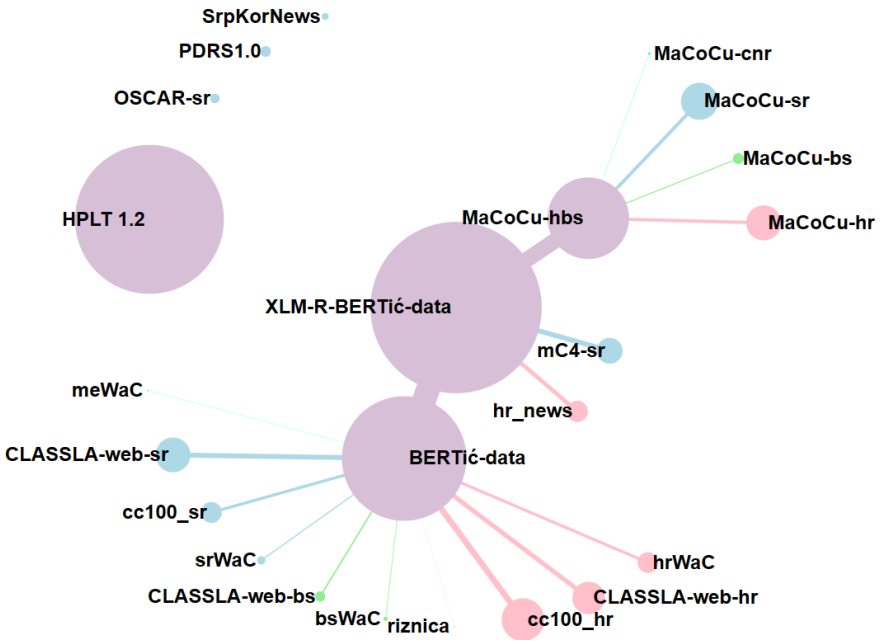
Још један значајан подухват дедупликације укључује четири *WaC* корпуса, три *CLASSLA* корпуса, *cc100_sr*, *cc100_hr* и корпус *riznica*. Тиме је произведен обједињени корпус од а 8,4 милијарди речи, објављен под именом *BERTić-data* (Ljubešić and Lauc 2021).

И *BERTić-data* и *MaCoCu-hbs* су поново обједињени, заједно са *mC4-sr* и *hr_news* (1,4 милијарде речи) како би се креирао *XLM-R-BERTić-data* (11,5 милијарди речи), највећи објављени српскохрватски обједињени корпус. Он, међутим, не укључује *Common Crawl* скупове *HPLT 1.2-sh* и *OSCAR-sr*, нити недавно објављени корпус *PDRS1.0* (Wasserscheidt 2023) (600M), а ни корпус *SrpKorNews* (Krstev and Stanković 2023) (468M). Ово указује на могућност креирања нових, већих кровних корпуса како за српски, тако и за српскохрватски језик. Комплетан садашњи подухват агрегације корпуса је приказан на слици 1.

Једини литерарни корпус доступан на вебу који испуњава величинске услове и даље је *SrpELTeC* корпус (Krstev 2021; Stanković et al. 2022), који обухвата 120 романа и садржи 5 милиона речи. Процес објављивања није био ограничен, јер корпус садржи материјале који су прекорачили праг истека ауторских права.

### 2.2 Јавно доступни обележени корпуси

Испитивањем одабраних чворишта са скуповима података пронашли смо 10 обележених корпуса за српски језик (табела 3). Усредсредили смо се на корпусе који су обележени помоћу скупа ознака *Universal POS* (17 класа), при чему није било ограничења када је у питању препознавање именованих ентитета (NER) и обележавање сентимента. За разлику од корпуса необележеног текста који су се примарно налазили на HF и VLO, ови корпуси су такође у значајном броју пронађени на ELG.




**Слика 1.** Хијерархија агрегације јавно доступних српскохрватских корпуса необележеног текста. Плава боја представља корпусе српског језика, цијан црногорског, ружичаста хрватског, зелена босанског, а љубичаста боја представља корпусе мешовитог језичког порекла.

Корпуси српског језика обележени према врсти речи укључују четири ресурса: *SrpKor4Tagging* (Stanković et al. 2020) са 60K, *Intera* (Gavrilidou et al. 2004; Stanković et al. 2020) са 47,5K, *1984* (Krstev, Vitas, and Erjavec 2004) са 6,7K и *reldi_sr* (Miličević and Ljubešić 2016) са 5,5K реченица. Два додатна ресурса у којима су обележене и врсти речи (POS) и именовани ентитети (NER): *SETimes_sr* (Samardžić et al. 2017) и *NormTagNER-sr* (Ljubešić et al. 2023) са 7K и 4K реченица, редом, и 5 NER класа. Три додатна ресурса у којима су обележени само именовани ентитети (NER): *SrpELTeC-gold* (Todorović et al. 2021; Frontini et al. 2020), заснован на делу литерарног корпуса који обухвата 52K реченица (7 NER класа), српски део корпуса *WikiAnn* (Rahimi, Li, and Cohn 2019) са 40K реченица (3 NER класе), и српски део корпуса *polyglot_ner* (Al-Rfou et al. 2015) са пола милиона реченица и

3 NER класе. Последњи ресурс је српски део *mms* корпуса обележених сентиментом (Augustyniak et al. 2023) са 76K реченица (3 класе).

| Назив | Јез. | Порекло | Анот. | Вел. | Издавач | Чвор. |
|---|---|---|---|---|---|---|
| SrpKor4Tagging | sr | mixed | POS | 60 | JeRTeH | ELG |
| 1984 | sr | literary | POS | 6,7 | MULTEXT-East | ELG |
| Intera | sr | textbook | POS | 47,5 | META-SHARE | ELG |
| reldi_sr | sr | web | POS | 5,5 | ReLDi | HF |
| SETimes_sr | sr | web | POS, NER | 4 | ReLDi | HF |
| NormTagNER-sr | sr | web | POS, NER | 7 | ReLDi | VLO |
| SrpELTeC-gold | sr | literary | NER | 52 | JeRTeh | ELG |
| wikiann-sr | sr | ∼textbook | NER | 40 | Wikimedia | HF |
| polyglot_ner-sr | sr | mixed | NER | 560 | Al-Rfou et al | HF |
| mms-sr | sr | web | sentiment | 76 | Brand24 | HF |

**Табела 3.** Обележени корпуси искључиво српског језика доступни на истраженим чвориштима скупова података. Величина је представљена у хиљадама реченица (K).

Слична ситуација је уочена када је у питању шира језичка слика, где је пронађено још десет ресурса (табела 4), при чему се листа већином састоји од хрватских еквивалената истих, претходно поменутих обележених ресурса. Ови ресурси обухватају обележени хрватски превод романа *1984*, потом *reldi_hr*, *NormTagNER-hr*, хрватски део *WikiAnn* и *polyglot_ner* корпуса и хрватске реченице обележене према сентименту из *mms* корпуса. Такође, ту је и босански *mms*, као и и српскохрватски *WikiAnn*, који је изведен из српскохрватских чланака на Википедији. Величине ових корпуса су углавном упоредиве са њиховим српским еквивалентима.

Једини јединствени ресурс са ове листе је *hr500k* (Ljubešić et al. 2016), којије обележен са NER ознакама (5 класа) у целости (25K реченица), а само делимично са скупом ознака Universal POS (9K реченица).

### 2.3 Јавно доступни паралелни корпуси

Укупно тринаест паралелних ресурса је пронађено за српски језик (табела 5), укључујући пет корпуса паралелних превода, један

| Назив | Јез. | Порекло | Анот. | Вел | Издавач | Чвор. |
|---|---|---|---|---|---|---|
| 1984 | hr | literary | POS | 6,7 | MULTEXT-East | ELG |
| hr500k | hr | literary | POS | 9 | ReLDi | HF |
| reldi_hr | hr | web | POS | 7 | ReLDi | HF |
| NormTagNER-hr | hr | web | POS, NER | 8 | ReLDi | VLO |
| hr500k | hr | literary | NER | 25 | ReLDi | HF |
| wikiann-hr | hr | ∼textbook | NER | 40 | Wikimedia | HF |
| wikiann-sh | sh | ∼textbook | NER | 40 | Wikimedia | HF |
| polyglot_ner-hr | hr | mixed | NER | 630 | Al-Rfou et al | HF |
| mms-hr | hr | web | sentiment | 78 | Brand24 | HF |
| mms-bs | bs | web | sentiment | 36 | Brand24 | HF |

**Табела 4.** Корпуси обележеног српскохрватског текста на истраженим чвориштима скупова података, који не садрже искључиво српски језик. Величина је представљена у хиљадама реченица (K).

корпус са паровима текст-резиме, један корпус парафраза, два корпуса са паровима питање-одговор (*QA*) и четири скупа података са инструкцијама (парови текстуалних инструкција и решења).

Сви *QA* корпуси и корпуси са инструкцијама су синтетичког порекла и креирани су помоћу машинског превођења постојећих корпуса на енглеском језику: *m_mmlu-sr* и *m_hellaswag-sr* су, редом, преводи *mmlu* (Hendrycks et al. 2021) и *HellaSwag* (Zellers et al. 2019) уз помоћ GPT3.5; *airoboros-3.0-sr* је превод скупа података *airoboros-3.0* (Tunstall et al. 2023), а *open-orca-slim-sr*, *ultrafeedback-bin-sr* и *alpaca-cleaned-sr* су српске верзије скупова података *SlimOrca* (Lian et al. 2023), *UltraFeedback* (Cui et al. 2023) и *alpaca-cleaned* (Taori et al. 2023), редом, преведени помоћу сервиса *Google translate*. Заједно са српским делом полусинтетичког скупа *tapaco* (Scherrer 2020) (оригинални текст можда није синтетички, али су реченице упариване коришћењем једноставног алгоритма), већина доступних паралелних ресурса за српски језик су синтетички.

Када су у питању ресурси који нису синтетички, пет корпуса паралелних превода (преводи са енглеског) обухватају литерарни корпус *80 jours* (Vitas et al. 2008) са 3,7K реченица и *biblenlp-sr*, који је изведен из електронске верзије Библије са 32K реченица derived from the elec-

tronic version of the *Bible* with 31K senteces, паралелну енглеско-српску верзију *Intera* корпуса (47,5K реченица), српско-енглески паралелни подскуп *Opus* скупа података (Tiedemann 2012), који обухвата 1 милион реченица, као и паралелизовани подскуп *MaCoCu* скупа података, *MacocuParallel-sr* (Banón et al. 2022), са 2 милиона реченица. Иако су два веб корпуса много већа по величини, могуће је да ће код њих бити неопходно уклањање дупликата.

Једини скуп података са сажецима је *XL-Sum* са 15 хиљада парова текстова и резимеа (Hasan et al. 2021).

| **Назив** | **Јез.** | **Порекло** | **Pair type** | **Вел.** | **Издавач** | **Чвор.** |
|---|---|---|---|---|---|---|
| 80 jours | sr | literary | translation | 3,7 | JeRTeh | ELG |
| biblenlp-sr | sr | literary | translation | 31 | eBible | HF |
| Intera | sr | mixed | translation | 47,5 | META-SHARE | ELG |
| OPUS-sr | sr | web | translation | 1000 | Opus project | HF |
| MacocuParallel-sr | sr | web | translation | 2000 | Bañón et al | HF |
| XL-Sum | sr | web | summary | 15 | Hasan et al | ELG |
| tapaco-sr | sr | ∼synthetic | paraphrase | 8 | Scherrer, Yves | HF |
| m_mmlu-sr | sr | synthetic | *QA* | 14,5 | Alexandra Inst. | HF |
| m_hellaswag-sr | sr | synthetic | *QA* | 9,5 | Alexandra Inst. | HF |
| airoboros-3.0-sr | sr | synthetic | instruction | 46 | draganjovanovich | HF |
| open-orca-slim-sr | sr | synthetic | instruction | 515 | DataTab | HF |
| ultrafeedback-bin-sr | sr | synthetic | instruction | 63 | DataTab | HF |
| alpaca-cleaned-sr | sr | synthetic | instruction | 52 | DataTab | HF |

**Табела 5.** Корпуси искључиво српског језика доступни на истраженим чвориштима скупова података. Величина је представљена у хиљадама речи(K).

Паралелни скупови података који нису на српском представљени су у табели 6 и састоје се већином од алтернативних подскупова истих ресурса: енглеско-хрватски подскуп скупа *biblenlp*, подскупови *OPUS* скупа који садрже преводе са српскохрватског, хрватског и босанског на енглески, као и подскупови скупа *MaCoCu Parallel*, који садрже преводе

са црногорског, хрватског и босанског на енглески. Ови скупови чине већину реченица, којих има преко 5 милиона.

Два јединствена ресурса са ове листе су хрватски подскуп *mfaq* скупа података који је сачињен од 5K парова питања и одговора сакупљених са веба (De Bruyn et al. 2021) и српскохрватски подскуп *exams* скупа података који садржи 5K парова питања и одговора преузетих са стварних тестова са ученицима (Hardalov et al. 2020). Упркос релативно малој величини, ови ресурси су важни у премало заступљеној групи *QA* скупова података, где су једина алтернатива синтетички преводи.

| **Назив** | **Јез.** | **Порекло** | **Врста пара** | **Вел.** | **Издавач** | **Чвор.** |
|---|---|---|---|---|---|---|
| biblenlp-hr | hr | literary | translation | 31 | eBible | HF |
| OPUS-sh | sh | web | translation | 271 | Opus project | HF |
| OPUS-hr | hr | web | translation | 1000 | Opus project | HF |
| OPUS-bs | bs | web | translation | 1000 | Opus project | HF |
| MacocuParallel-me | cnr | web | translation | 218 | Bañón et al | HF |
| MacocuParallel-hr | hr | web | translation | 2000 | Bañón et al | HF |
| MacocuParallel-bs | bs | web | translation | 500 | Bañón et al | HF |
| mfaq-hr | hr | web | *QA* | 5 | CLiPS | HF |
| exams-sh | mixed | textbook | *QA* | 5 | Hardalov et al | HF |

**Табела 6.** Српскохрватски паралелни корпуси доступни на истраженим чвориштима скупова података, који нису искључиво на српском. Величина је приказана у хиљадама реченица (K).

## 3. Нови корпуси

У оквиру овог истраживања предвиђена су три нова корпуса. Први (*Кишобран*) је нови и већи кровни корпус необележеног текста који ће који ће обухватити све тренутно објављене веб корпусе под једним ентитетом. Израда корпуса Кишобран је описана у одељку 3.1. Други предвиђени корпус (*S.T.A.R.S.*) осмишљен је да повећа заступљеност јавно доступних извора високог квалитеета и заснован је на докторским дисертацијама преузетим са платформе НаРДуС,[7] које су претходно

7. НаРДуС – Национални репозиторијум дисертација у Србији.

коришћене за тренирање тренутно највећег језичког модела који је претрениран за српски језик, *gpt2-orao* (Škorić 2024). Процес припреме *S.T.A.R.S.* корпуса биће објашњен детаљније у одељку 3.2. Последњи корпус (*PaSaž*) обухвата поравнате преводе сажетака издвојених из ових дисертација помоћу метода представљених у одељку 3.3.

### 3.1 Кровни веб корпус - Кишобран

Нови српскохрватски кровни корпус, *Umbrella corp.* резултат је обједињавања двадесет постојећих корпуса који су прегледани и разматрани у овом раду. Избегавали смо коришћење постојећих кровних корпуса као материјала како бисмо обезбедили могућност касније екстракције одређеног језика, на пример српског дела корпуса, али треба напоменути да је већина коришћених корпуса већ била дедупликована како у односу на друге корпусе са листе, тако и интерно. Једини корпус са мешаним језицима на листи, *HPLT 1.2-sh* је додатно обрађен како би био подељен у корпусе који садрже документа на српском (*HPLT-sr*) и хрватском језику (*HPLT-hr*). Основа за ову врсту меке класификације је била учесталост специфичних речи (тко/ко, што/шта, увјет/услов, уопће/уопште, итд.), као и однос слова *е* и подниске *је*, указујући на ијекавски дијалекат.

Корпуси су додатно пречишћени и дедупликовани. Дедупликација садржаја корпуса је извршена на свим документима помоћу *onion* (Pomikálek 2011) алата за обраду корпуса који врши расплинуту дедупликацију и лоцира дупликате докумената на основу н-торки дуплираних кроз читав корпус. За овај експеримент користили смо претрагу дупликата шесторки и елиминисали све документе изнад прага дуплицирања од 75%.

Након ове дедупликације укупан број речи је смањен за трећину, са 28 милијарди речи на 18,6 милијарди. Са овим укупним бројем, Кишобран је постао највећи доступни кровни корпус необележеног текста уједичком опсегу, са додатним побољшањима (као што су додатно уклањање артефакта и корекција текста) планираним у будућности. .

Целокупна листа коришћених корпуса, њихове величине у милионима речи пре и после дедупликације и чишћења, као и њихов укупни удео у финалној верзији корпуса *Umbrella corp.* представљени су у табели 7.

| Име | Јез. | Вел. пре | Вел. после | Удео |
|---|---|---|---|---|
| srWaC | Serbian | 493 | 307 | 1,65% |
| meWaC | Montenegrin | 80 | 41 | 0,22% |
| hrWaC | Croatian | 1.250 | 935 | 5,01% |
| bsWaC | Bosnian | 256 | 194 | 1,04% |
| OSCAR-sr | Serbian | 632 | 410 | 2,20% |
| cc100_hr | Croatian | 2.880 | 2,561 | 13,73% |
| cc100_sr | Serbian | 711 | 659 | 3,53% |
| CLASSLA-sr | Serbian | 752 | 240 | 1,29% |
| CLASSLA-hr | Croatian | 1.341 | 160 | 0,86% |
| CLASSLA-bs | Bosnian | 534 | 105 | 0,56% |
| hr_news | Croatian | 1.433 | 1.426 | 7,65% |
| mC4-sr | Serbian | 800 | 782 | 4,19% |
| MaCoCu-sr | Serbian | 2.491 | 2.152 | 11,54% |
| MaCoCu-cnr | Montenegrin | 161 | 152 | 0,82% |
| MaCoCu-hr | Croatian | 2.363 | 2.355 | 12,63% |
| MaCoCu-bs | Bosnian | 730 | 700 | 3,75% |
| riznica | Croatian | 87 | 69 | 0,37% |
| PDRS1.0 | Serbian | 602 | 506 | 2,71% |
| SrpKorNews | Serbian | 469 | 469 | 2,51% |
| HPLT-sr | Serbian | ∼5.015 | 2.562 | 9,95% |
| HPLT-hr | Croatian | ∼5.015 | 1.856 | 13,74% |
| Total | | 28,095 | **18.641** | 100% |

**Табела 7.** Листа корпуса коришћених за израду новог кровни корпуса, њихове величине пре и после чишћења и дедупликације и њихов удео у финалном корпусу. Величине су представљене у милионима речи (M).

### 3.2 Скуп теза и академских радова на српском – *S.T.A.R.S.*

У домену корпуса високог квалитета ($O_2$), докторске тезе представљају веома пожељан тип ресурса због неколико фактора. Најпре, високи академски стандарди у погледу методологије, података и језичког квалитета које докторска теза мора да задовољи, као и процес рецензије кроз који пролази, обезбеђују поуздан висок

квалитет података из овог типа извора. Затим, пошто се од докторске дисертације очекује да представља јединствен научни допринос који ће додатно унапредити релевантну научну област, корпус докторских дисертација заузима јединствену позицију када су у питању тематска разноликост у различитим областима знања и актуелност описаних тема. Коначно, велики број докторских дисертација у отвореном приступу у НаРДуС репозиторијуму, чињеница да су одређени одељци докторских дисертација стандардизовани, као и доступност структурираних метаподатака за сваку дисертацију на НаРДуС-у, чине овај ресурс идеалним кандидатом за јединствен, велики и висококвалитетни научни корпус на српском.

НаРДуС је осмишљен у оквиру структурног ТЕМПУС пројекта 5440932013, RODOS.[8] На основу измена и допуна Закона о Високом образовању[9] („Службени гласник РС“, бр. 99/2014, ступивши на снагу дана 19. 9. 2014), све високообразовне институције имају обавезу да учине докторске дисертације доступним јавности пре њихове одбране, док универзитети имају обавезу да обезбеде дигитални репозиторијум са отвореним приступом одбрањеним дисертацијама. У складу са наведеним, платформа НаРДуС је у функцији од септембра 2015. године и заснована је на софтверу *DSpace*, који подржава *OAI-PMH* протокол за пренос метаподатака (Verbić, Suvakov, and Luzanin 2017). У тренутку писања овог рада, на платформи НаРДуС се налазе 13.289 докторских дисертација.

Као први корак у креирању корпуса, комплетна листа дисертација је преузета са мапе сајта платформе НаРДуС,[10] уз помоћ Пајтон модула *Requests*, поштујући спецификације странице *robots.txt* на овом сајту. За преузимање комплетних детаљних метаподатака сваке дисертације, укључујући линкове за преузимање, коришћени су Пајтон и библиотека *Selenium*. Метаподаци за сваку дисертацију су обогаћени са четири додатна поља:

***fulltext_url*** – У случајевима када је за једну дисертацију било доступно више линкова за преузимање (у које је често био укључен извештај комисије), сви PDF документи са линкова су преузети, број страница и линија је аутоматски добијен коришћењем модула

8. rodos.edu.rs Restructuring of Doctoral Studies in Serbia (RODOS)
9. www.pravno-informacioni-sistem.rs/SlGlasnikPortal/reg/viewAct/5f688c8e-4798-470f-9adf-5bc0739c72aa
10. nardus.mpn.gov.rs/htmlmap

*PyMuPDF* у Пајтону, а одговарајући линк је одабран и сачуван у овом пољу;

***need_ocr*** – Покушано је аутоматско преузимање текста помоћу модула *PyMuPDF* за првих 10 страница сваке дисертације, да би се проценило да ли су документи са одабраних URL-ова дигитално настали или захтевају даље кораке оптичког препознавања знакова (OCR) и ручну ревизију. Резултати ове процедуре су сачувани у овом пољу у виду буловских вредности;

***srpski*** – Ово поље приказује да ли су дисертације написане на српском језику (вредност *yes*) или на неком другом језику (вредност *no*), а језик сваког од тектова је одређиван испитивањем поља *dc.language* и *dc.language.iso*;

***ARR*** – Ово поље приказује да ли су дисертације објављене под заштићеном лиценцом – *all rights reserved* (ARR), што је утврђено прегледом поља *dc.rights.license* у метаподацима.

Ова четири поља су коришћена за филтрирање одговарајућих дисертација-кандидата за ову верзију корпуса. Изабрали смо оне где је било могуће добити се одговарајући PDF могао добити помоћу поља *fulltext_url*, који су написани на српском (*srpski=yes*), који не захтевају OCR (*ocr=no*), и који нису објављени под лиценцом *all rights reserved*. Ово филтрирање је урађено да би се издвојили само текстови на српском језику и како би се осигурала усклађеност са лиценцама ауторских права.

Након примене овог критеријума, остало је 11.624 дисертације, што је представљало око 87,5% свих дисертација у НаРДуС-у.

На крају, цео текст из сваке од ових дисертација је издвојен уз помоћ модула *PyMuPDF* и сачуван у облику текстуалних датотека, које су коришћене за изградњу корпуса. Само линије које су биле део пасуса текста су задржане у састављању финалног корпуса, што је резултирало корпусом са преко 560 милиона речи.

### 3.3 Корпус паралелних сажетака – *ПаСаж*

С обзиром на то да су паралелни језички ресурси, посебно за мање распрострањене језике (као што је српски), релативно ретки, паралелни корпуси су мањи, али и мање бројни у односу на једнојезичне корпусе. Пошто су сажеци српских докторских дисертација написани на српском и енглеском језику, они представљају вредан велики и квалитетан ресурс за креирање паралелног српско-енглеског корпуса

научног језика. Међутим, наслов, формат и локација сажетака унутар документа значајно варирају у зависности од научне институције. У овој студији прво је извршена ручна анализа великог броја дисертација из различитих институција, након чега су документи подељени у две групе; у првој су биле дисертације чији су се сажеци налазили у самом тексту (било на почетку или крају документа), а у другој дисертације чији су сажеци били представљени у оквиру табеле у одељку *Кључна документацијска информација*. Од 11.624 дисертације, 2.594 су другој групи, која је користила табеле. Из докумената су извучене три врсте података (како на српском, тако и на енглеском): сажеци, кључне речи и научна област дисертације. Пошто су кључне речи већ присутне у метаподацима, њихово извлачење је рађено само за прву групу докумената, где су постојала два скупа кључних речи (због могућих разлика између две верзије). За другу групу су извучени само сажеци и научна област дисертације.

С обзиром на то да метаподаци за већину дисертација у НаРДуС-у садрже делимичне сажетке на оба језика, Пајтон скрипта је коришћена за налажење почетка оба сажетка (првих 6 речи) у првој групи докумената. Уколико дисертација није имала делимичан сажетак у метаподацима или је претрага била неуспешна, налажење почетка сажетка је покушано помоћу регуларних израза који су одговарали могућем наслову одељка сажетка. Пошто су информације у овом одељку увек биле представљене следећим редом: 1) сажетак, 2) кључне речи, 3) научна област (за оба језика), примењен је приступ који је користио регуларне изразе за идентификовање краја сваког сегмента и почетка наредног, након чега су са овим информацијама ажурирани постојећи метаподаци дисертације.

За другу групу, у којој су жељени одељци били присутни у делу *Кључна документацијска информација*, исти приступ је примењен, уз коришћење другачијих регуларних израза за сегменте табела.

Након ових корака, извучено је 7.687 паралелних сажетака, а метаподаци ових дисертација су ажурирани сажецима и научном облашћу дисертације. Где је било могуће, кључне речи из текста дисертација су такође извучене и сачуване у новом пољу у метаподацима (*keywords_from_text*), у случају да се вредности разликују од оних већ постојећих у метаподацима добијеним са НаРДуС-а.

## 4. Евалуација

Евалуација података у оквиру овог рада фокусира се само на корпусе необележеног текста на српском језику, а сама евалуација се врши кроз грубу процену јединствености сваког корпуса, односно процене колико се он разликује од осталих. Да бисмо проценили аспект јединствености, одлучили смо да спроведемо евалуацију засновану на учесталости речи, инспирисану постојећим експериментом (Rayson and Garside 2000). За сваки корпус-кандидат, извод од милион речи је коришћен за компилацију фреквенција речи, а затим је извучено 1000 најчешћих речи из сваког извода корпуса. Ове најчешће речи из сваког од десет извода корпуса коришћене су за изградњу листе карактеристика (јединственог скупа речи), што је резултирало са укупно 3.257 карактеристика из десет корпуса. Релативне фреквенције (по милион речи) сваке речи из скупа карактеристика у сваком корпусу коришћене су за попуњавање вектора карактеристика (ако је реч једна од одабраних 1000 за тај корпус, у супротном је релативној фреквенцији додељена вредност 0). Када су вектори карактеристика попуњени, израчунате су косинусне сличности (на десети) између сваког пара како би се генерисала матрица сличности корпуса. Корпус који има најмању релативну сличност са осталима сматра се најјединственијим. Израчуната матрица сличности корпуса приказана је у табели 8.

Из представљених резултата је евидентно да је најјединственији корпус према учесталости речи *SrpELTeC*, са просечном сличношћу од 0,40, посебно у поређењу са укупном просечном сличношћу од 0,71. Такође је најудаљенији од сваког појединачног корпуса у матрици сличности.

Нови корпус, *S.T.A.R.S.*, је други по удаљености од свих осталих корпуса појединачно и у просеку, што га чини другим најјединственијим корпусом (просечна сличност од 0,57). Оно што је посебно занимљиво јесте да су два корпуса са најмањом међусобном сличношћу управо *S.T.A.R.S.* и *SrpELTeC*, са сличношћу од само 0,22, што их ставља на супротне крајеве спектра, док се веб корпуси групишу у средини, што је посебно видљиво на дводимензионалној репрезентацији матрице сличности. (слика 2).

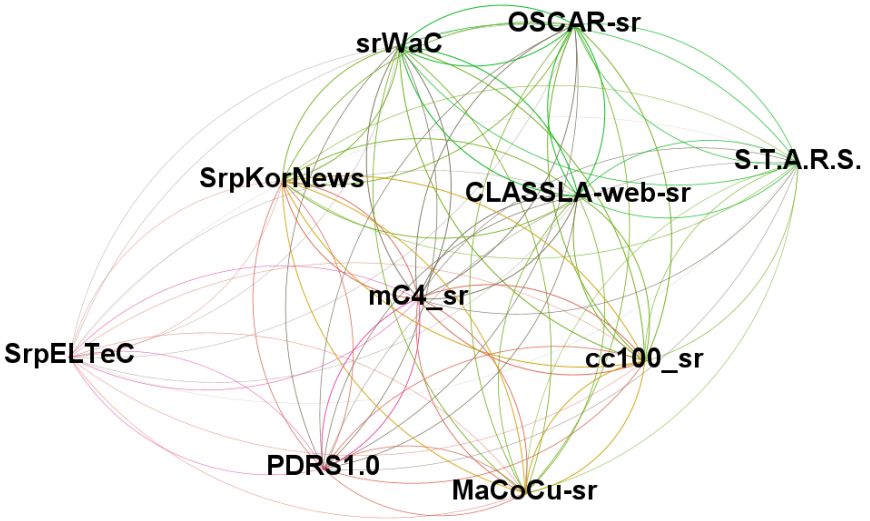
| | srWaC | cc100_sr | mC4_sr | OSCAR-sr | CLASSLA-sr | MaCoCu-sr | PDRS1.0 | SrpKorNews | SrpELTeC | S.T.A.R.S. |
|---|---|---|---|---|---|---|---|---|---|---|
| srWaC | | 0,93 | 0,88 | 0,95 | 0,98 | 0,79 | 0,72 | 0,87 | 0,36 | 0,71 |
| cc100_sr | 0,93 | | 0,91 | 0,91 | 0,90 | 0,92 | 0,75 | 0,93 | 0,44 | 0,62 |
| mC4_sr | 0,88 | 0,91 | | 0,84 | 0,87 | 0,84 | 0,78 | 0,88 | 0,50 | 0,61 |
| OSCAR-sr | 0,95 | 0,91 | 0,84 | | 0,94 | 0,78 | 0,70 | 0,82 | 0,37 | 0,69 |
| CLASSLA-sr | 0,98 | 0,90 | 0,87 | 0,94 | | 0,75 | 0,71 | 0,85 | 0,34 | 0,70 |
| MaCoCu-sr | 0,79 | 0,92 | 0,84 | 0,78 | 0,75 | | 0,71 | 0,89 | 0,48 | 0,52 |
| PDRS1,0 | 0,72 | 0,75 | 0,78 | 0,70 | 0,71 | 0,71 | | 0,73 | 0,47 | 0,47 |
| SrpKorNews | 0,87 | 0,93 | 0,88 | 0,82 | 0,85 | 0,89 | 0,73 | | 0,42 | 0,56 |
| SrpELTeC | 0,36 | 0,44 | 0,50 | 0,37 | 0,34 | 0,48 | 0,47 | 0,42 | | 0,22 |
| S,T,A,R,S, | 0,71 | 0,62 | 0,61 | 0,69 | 0,70 | 0,52 | 0,47 | 0,56 | 0,22 | |
| *Average* | 0,80 | 0,81 | 0,79 | 0,78 | 0,78 | 0,74 | 0,67 | 0,77 | 0,40 | 0,57 |

**Табела 8.** Матрица сличности српских корпуса необележеног текста доступних на испитиваним чвориштима скупова података, заснована на учесталости речи. Вредности представљају косинусне сличности подигнуте на десети степен, добијене поређењем вектора учесталости речи из скупова карактеристика.

## 5. Закључак

Овај рад пружа преглед текстуалних корпуса за српски (и српскохрватски) језик који су јавно доступни (на чвориштима *Hugging Face, European Language Grid* и *CLARIN VLO*) у следећим категоријама:

– Корпуси необележеног текста – углавном веб порекла, са изузетком књижевног корпуса *SrpELTeC*;
– Обележени корпуси – различитог порекла, примарно обележени са POS и NER информацијама;
– Паралелни корпуси – углавном веб порекла (*MacocuParallel*, *OPUS*), књижевни преводи и синтетички *QA* сетови и сетови инструкција, један веб корпус резимеа, један полусинтетички корпус парафраза и два мања уређена *QA* скупа.




**Слика 2.** Јединственост корпуса визуализована кроз дводимензионалну репрезентацију матрице сличности корпуса у виду графа мреже, где ивице представљају удаљености између сваког корпуса, а боје представљају спектар груписања.

Утврдили смо да велика већина доступних корпуса необележеног текста потиче са веба, и да постоји више напора њихове дедупликације, али ниједан тренутно не обухвата све корпусе. Такође смо утврдили да су корпуси високог квалитета веома слабо заступљени, без ијединог корпуса који испуњава критеријум величине од најмање три милиона речи.

Да бисмо превазишли недостатак текстова високог квалитета, предлажемо и састављамо два нова, високо уређена корпуса на основу јавно доступних докторских дисертација на српском: корпус необележеног текста назван *S.T.A.R.S.* (Set of theses and academic research in Serbian) и и паралелни енглеско-српски корпус сажетака назван *PaSaž*. Први садржи 11.624 документа и преко 560 милиона речи, а други садржи 7.678 паралелних сажетака и додатних 2.805 делимичних сажетака, што чини око осам милиона речи, и представља велики допринос јавно доступним корпусима високог квалитета за моделовање српског језика.

Такође смо спровели евалуацију засновану на фреквенцији речи која указује на јединственост дисертација у тренутној парадигми, где је показано да имају релативно ниску сличност са другим доступним корпусима, нарочито са књижевним корпусом *SrpELTeC*, док се веб корпуси (који имају високу међусобну сличност) налазе у средини (слика 2).

Поред тога, рад представља нови кровни веб корпус за српски и српскохрватски језик назван *Umbrella corp.*, који обједињује све тренутно доступне веб корпусе необележеног текста у планираном језичком обиму.

Будући рад у овој области треба да укључи даље прибављање висококвалитетних и књижевних текстова који се могу објавити, као и креирање процедура за унапређење постојећих ресурса, посебно веб текстова.

## Захвалница

Рачунарски ресурси неопходни за дедупликацију корпуса *Umbrella corp.* обезбеђени су од стране Националне платформе за вештачку интелигенцију Србије.

Ово истраживање је подржано од стране Фонда за науку Републике Србије, #7276, *Text Embeddings – Serbian Language Applications – TESLA*.

## Литература

Augustyniak, Łukasz, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. 2023. *Massively Multilingual Corpus of Sentiment Datasets and Multi-faceted Sentiment Classification Benchmark.* arXiv: 2306.07902 [cs.CL].

Aulamo, Mikko, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer Van Der Linde, and Jaume Zaragoza. 2023. "HPLT: High Performance Language Technologies." In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation,* 517–518.

Banón, Marta, Miquel Espla-Gomis, Mikel L Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, et al. 2022. "MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages." In *23rd Annual Conference of the European Association for Machine Translation, EAMT 2022,* 303–304. European Association for Machine Translation.

Ćavar, Damir, and Dunja Brozović Rončević. 2012. "Riznica: the Croatian language corpus." *Prace filologiczne* 63:51–65.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "Unsupervised cross-lingual representation learning at scale."

Cui, Ganqu, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. *UltraFeedback: Boosting Language Models with High-quality Feedback.* arXiv: 2310.01377 [cs.CL].

De Bruyn, Maxime, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. *MFAQ: a Multilingual FAQ Dataset.* arXiv: 2109.12870 [cs.CL].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805.*

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020.*

Gavrilidou, Maria, Penny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos, and Stelios Piperidis. 2004. "Building parallel corpora for eContent professionals." In *Proceedings of the Workshop on Multilingual Linguistic Resources,* 90–93.

Hardalov, Momchil, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. "EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),* edited by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, 5427–5444. Online: Association for Computational Linguistics, November. https://doi.org/10.18653/v1/2020.emnlp-main.438. https://aclanthology.org/2020.emnlp-main.438.

Hasan, Tahmid, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. "XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages." In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021,* 4693–4703. Online: Association for Computational Linguistics, August. https://aclanthology.org/2021.findings-acl.413.

Hendrycks, Dan, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. "Aligning AI With Shared Human Values." *Proceedings of the International Conference on Learning Representations (ICLR).*

Krstev, Cvetana. 2021. "The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata." *Infotheca - Journal for Digital Humanities* 21 (2): 26–42. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.2. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.2_en.

Krstev, Cvetana, and Ranka Stanković. 2023. "Language Report Serbian." In *European Language Equality: A Strategic Agenda for Digital Language Equality,* edited by Georg Rehm and Andy Way, 203–206. Cham: Springer International Publishing. ISBN: 978-3-031-28819-7. https://doi.org/10.1007/978-3-031-28819-7_32.

Krstev, Cvetana, Duško Vitas, and Tomaž Erjavec. 2004. "MULTEXT-East resources for Serbian." In *Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 Jezikovne tehnologije 9-15 Oktober 2004, Ljubljana, Slovenija, 2004.* Erjavec, Tomaž and Zganec Gros, Jerneja.

Kuzman, Taja, and Nikola Ljubešić. 2023. *Montenegrin web corpus meWaC 1.0.* CLASSLA-web: Bigger and better web corpora for Croatian, Serbian and Slovenian. https://www.clarin.si/info/k-centre/classla-web-bigger-and-better-web-corpora-for-croatian-serbian-and-slovenian-on-clarin-si-concordancers/.

Lian, Wing, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. *SlimOrca: An Open Dataset of GPT-4 Augmented FLAN Reasoning Traces, with Verification.* https://https://huggingface.co/Open-Orca/SlimOrca.

Ljubešić, Nikola, and Tomaž Erjavec. 2021. *Montenegrin web corpus meWaC 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1429.

Ljubešić, Nikola, Tomaž Erjavec, Vuk Batanović, Maja Miličević, and Tanja Samardžić. 2023. *Serbian Twitter training corpus ReLDI-NormTagNER-sr 3.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1794.

Ljubešić, Nikola, and Filip Klubička. 2014. "{bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian." In *Proceedings of the 9th web as corpus workshop (WaC-9),* 29–35.

Ljubešić, Nikola, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* edited by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, et al. Portorož, Slovenia: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-9-1.

Ljubešić, Nikola, and Davor Lauc. 2021. "BERTić–The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian." *arXiv preprint arXiv:2104.09243.*

Miličević, Maja, and Nikola Ljubešić. 2016. "Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets." *Slovenščina 2.0: empirical, applied and interdisciplinary research* 4, no. 2 (September): 156–188. https://doi.org/10.4312/slo2.0.2016.2.156-188. https://revije.ff.uni-lj.si/slovenscina2/article/view/7007.

Pomikálek, Jan. 2011. "Removing boilerplate and duplicate content from web corpora." PhD diss., Masaryk university, Faculty of informatics, Brno, Czech Republic.

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. "Improving language understanding by generative pre-training."

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21 (1): 5485–5551.

Rahimi, Afshin, Yuan Li, and Trevor Cohn. 2019. "Massively Multilingual Transfer for NER." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 151–164. Florence, Italy: Association for Computational Linguistics, July. https://www.aclweb.org/anthology/P19-1015.

Rayson, Paul, and Roger Garside. 2000. "Comparing corpora using frequency profiling." In *The workshop on comparing corpora,* 1–6.

Al-Rfou, Rami, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. "Polyglot-NER: Massive Multilingual Named Entity Recognition." *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30- May 2, 2015* (April).

Samardžić, Tanja, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. "Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages." In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing,* 39–44. Valencia, Spain: Association for Computational Linguistics, April. https://doi.org/10.18653/v1/W17-1407. https://aclanthology.org/W17-1407.

Scherrer, Yves. 2020. *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages.* V. 1.0, March. https://doi.org/10.5281/zenodo.3707949. https://doi.org/10.5281/zenodo.3707949.

Škorić, Mihailo. 2024. "Novi jezički modeli za srpski jezik." Accepted for publishing, *Infoteka* 24 (1).

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, and Milica Ikonić Nešić. 2022. "Distant Reading in Digital Humanities: Case Study on the Serbian Part of the ELTeC Collection." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 3337–3345.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian" [in English]. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, et al., 3954–3962. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.487.

Suárez, Pedro Javier Ortiz, Benoît Sagot, and Laurent Romary. 2019. "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures." In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7).* Leibniz-Institut für Deutsche Sprache.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. *Stanford Alpaca: An Instruction-following LLaMA model.* https://github.com/tatsu-lab/stanford_alpaca.

Tiedemann, Jörg. 2012. "Parallel Data, Tools and Interfaces in OPUS." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12),* edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2214–2218. Istanbul, Turkey: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Todorović, Branislava Šandrih, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian ner&beyond: The archaic and the modern intertwinned." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260.

Tunstall, Lewis, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, et al. 2023. *Zephyr: Direct Distillation of LM Alignment.* arXiv: 2310.16944 [cs.LG].

Verbić, Srđan, Milovan Suvakov, and Zorana Luzanin. 2017. "NaRDuS – JAVNOST DOKTORSKIH STUDIJA Transparentnost i otvorenost podataka kroz stvaranje i razvoj nacionalnog repozitorijuma doktorskih disertacija u Srbiji (NaRDuS)." June.

Vitas, Duško, Svetla Koeva, Cvetana Krstev, and Ivan Obradović. 2008. "Tour du monde through the dictionaries." In *Actes du 27eme Colloque International sur le Lexique et la Gammaire,* 249–256.

Vitas, Duško, Ljubomir Popović, Cvetana Krstev, Ivan Obradović, Gordana Pavlović-Lažetić, and Mladen Stanojević. 2012. *Српски језик у дигиталном добу – The Serbian Language in the Digital Age.* META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Available online at http://www.meta-net.eu/whitepapers. Springer. ISBN: 978-3-642-30754-6.

Wasserscheidt, Philipp. 2023. *Serbian Web Corpus PDRS 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1752.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, 483–498. Online: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/2021.naacl-main.41. https://aclanthology.org/2021.naacl-main.41.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. "HellaSwag: Can a Machine Really Finish Your Sentence?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.*

# Рачунар „галаксија“ – пут до звезда поплочан речима

УДК 050:004GALAKSIJA"1983/1984"

**САЖЕТАК:** Истраживање је конципирано као анализа садржаја научнопопуларног часописа *Галаксија* (бројева који су излазили током 1983. и почетком 1984. године) са циљем да се утврди како се у одређеним текстовима приказују рачунар, робот и сам концепт вештачке интелигенције (AI – Artificial Intelligence), као и у којој мери је такво писање допринело великом успеху рачунара „галаксија“, првог југословенског рачунара намењеног широкој употреби. У раду се истиче да се решавању проблема приступило на креативан и довитљив начин у тренуцима када су постојала објективна ограничења која је наметнула држава.

**КЉУЧНЕ РЕЧИ:** галаксија, рачунар, вештачка интелигенција, анализа садржаја.

**РАД ПРИМЉЕН:** 27. јануар 2024.
**РАД ПРИХВАЋЕН:** 15. март 2024.

Олга Мирковић Максимовић
olga.mirkovic@gmail.com
ORCID: 0000-0003-4255-9200
*Универзитет у Београду,*
*Мултидисциплинарне докторске студије*
*Историја и филозофија природних наука и технологије*
*Београд, Србија*

## 1. Рачунар „галаксија“ – око чега сва та бука?

Микрорачунар „галаксија“ је први домаћи рачунар доступан најширем кругу обожавалаца у периоду интензивног развоја рачунара и њиховог приближавања крајњим корисницима, како у свету, тако и на просторима бивше СФРЈ. Важно је разумети да током 1983. године увоз рачунара у тадашњу Југославију био забрањен, те је љубитељима рачунара и технике било немогуће да дођу до жељеног предмета увозом на легалан начин. Са циљем да се домаћа производња заштити од стране конкуренције, у СФРЈ су уведена строга царинска правила за увоз електронске опреме, што се одразило и на рачунаре. Влада је покушала да контролише прилив страних кућних рачунара

уводећи увозна ограничења по питању цена и величине меморије. Како је жеља била јача од ограничења, људи су се досетили да уносе и увозе део по део рачунара и састављају га не кршећи тако важећа правила и прописе. Људи су се довијали и уз помоћ пријатеља и рођака из иностранства легално набављали најразличитије делове. Притисци државе само су поспешили креативност и инвентивност љубитеља рачунара, и подстакли њихово самоорганизовање у жељи да коначно буду власници рачунара погодног за кућну употребу. Најбољу илустрацију ширења микрорачунара проналазимо у 129. броју *Галаксије* у оквиру текста „Микрорачунар изблиза“, где се каже: „Милиони људи најразличитијих занимања, ђака и студената, песника и домаћица, радника и службеника, лекара и адвоката већ су у најближем контакту са компјутером.“

Више компанија покушало је да направи микрорачунаре сличне кућним рачунарима из 1980-их, на пример Институт „Иво Лола Рибар“ – „лола 8“, ЕИ Ниш – „ресот“ 32 и 64, ПЕЛ Вараждин – „галеб“ и „орао“, Ивасим „ивел ултра“ и „ивел Z3“ итд. (Le Parisien 2000). На њихов неуспех или позиционирање ван тржишта кућних рачунара утицали су многи фактори:

– били су прескупи за појединачне купце (посебно у поређењу са популарним страним рачунарима „ZX Spectrum“, „Commodore 64“ итд.),
– мали избор забавних и осталих програма није привлачио рачунарске ентузијасте,
– није било могуће купити их у продавницама.

Домаћи рачунари ове генерације углавном су се користили у државним институцијама јер је и њима било забрањено да купују увозне рачунаре.

Рачунар „галаксија“ био је једини који је успео да се избори за своје место под сунцем и освоји срца љубитеља кућних рачунара у СФРЈ. Његов творац је Воја Антонић. У интервјуу који је дао за портал „Стартит“ Воја прича када и како је дошао на идеју да га направи:

„(идеја)... Настала је у Црној Гори на летовању. За мене су летовања на мору бескрајно досадна јер нема шта да се ради. Тамо сам узео свеску и на плажи почео да правим неке своје пројекте и цртам. Ту сам први пут дошао на идеју како може једноставно и економично да се направи видео-јединица, која је била најкомпликованија ствар за тадашње компјутере. Цена компјутера је увек зависила од сложености

видео-јединице. Ја сам дошао на замисао како то може врло једноставно да се реши, да се направи компјутер који је лош компјутер, спор до бестрага јер четири петине времена троши на генерисање слике, али је то ипак био компјутер. Схватио сам да такав рачунар ради и може врло једноставно да се направи, па сам помислио *зашто да не понудим још некоме*? Онда сам решио да ћу то негде да објавим као „do it yourself“ (уради сам) пројекат. Часопис *Галаксија* је објављивао специјално издање о рачунарима које је требало да се зове „Рачунари у вашој кући“. Повезао сам се са Дејаном Ристановићем, човеком који је требало то да пише, и испричао му своју идеју, на шта је он рекао: 'Како да не. То би било дивно да за први број буде нека самоградња' и тако ми кренемо“ (Startit 2017).

У истом интервјуу сазнајемо и каква су била очекивања, колико ће људи желети да састави „галаксију“, и она су се кретала од 50 до 500. Број наруџбеница пристиглих у часопис *Галаксију* за склапање рачунара „галаксија“ износио је 8.000 и превазишао је и најсмелија надања.

У интервјуу за *Југопапир* о успеху акције сведочи и други актер догађаја – Дејан Ристановић, уредник специјалног издања *Галаксије* под називом „Рачунари у вашој кући“. Ристановић говори о томе како су погрешили у процени тиража.

„Први број 'Рачунара у вашој кући' штампали смо у 30.000 примерака. Распродати су за две недеље, тако да смо морали да доштампамо 20.000. И то смо продали, тако да је на крају направљено поновно издање првог броја у још 25.000 примерака. Други је број штампан одмах у 50.000 примерака, а како је распродат, мислим да ће се ових дана доштампати 25.000 примерака. Трећи број се прави у 50.000 примерака. Да смо ствари мало боље проценили и штампали од прве по 100.000 комада, опет би све било продато јер када више није било Рачунара на киосцима, људи су их размењивали и фотокопирали“ (Yugopapir 2014).

О томе како су се спојили часопис *Галаксија* и истоимени рачунар такође сведочи Ристановић на свом блогу: „Да би оваква акција успела, требало је неколико месеци припремати и пратити кроз текстове у *Галаксији*, па је тако септембра 1983. објављен мој текст у коме је самоградња представљена. Текст је пратила прелиминарна наруџбеница, а одзив је био невероватан: преко хиљаду читалаца *Галаксије* изражава жељу да сагради свој први рачунар! Овако велики одзив је вероватно последица чињенице да је увоз рачунара (као и било које друге робе скупље од 1.500 тадашњих динара) био забрањен и да хиљаде људи

заинтересованих за рачунаре није видело никакав други начин да прошири своје знање. Карактеристике рачунара „галаксија“ нису биле богзна како добре, али је наш првенац био употребљив, уравнотежен и, што је најважније, достижан – требало је само скупити храброст и узети лемилицу у руке!“ (Ristanović).

Техничке карактеристике рачунара „галаксија“ (слика 1) биле су следеће:

- Процесор: ZiLOG Z80A на 3,072 MHz
- Галаксија ROM „A“ (ROM „A“ или „1“) – 4 kB EPROM (2732 EPROM) са једноставним оперативним системом, и Галаксија BASIC код за BASIC програм-преводилац;
- Галаксија ROM „B“ (ROM „B“ или „2“) – 4 kB (опционо, исто 2.732 EPROM), даје додатне BASIC команде, машински језик (асемблер), монитор машинског кода и друго;
- Галаксија ROM за знакове – 2 kB (2716 EPROM) садржи дефиниције симбола (слова и знакова);
- Меморија RAM: од 2 до 6 kB статичке меморије (6116 RAM) у основној верзији, могуће проширење до 54 килобајта;
- Текст: 32 × 16 слова или знакова, црно-бело;
- Псеудографика: 2 × 3 матрица, укупно 64 × 48 тачака;
- Звук: Нема, али интерфејс за касетофон је могао да се користи у ту сврху, као на „спектруму“ (ZX Spectrum);
- Спремање података: касетофонска трака, снимање брзином од 280 бита у секунди;
- Улазни и излазни портови: 44-пински ивични конектор („Edge connector“) са Z80 бусом, касетни магнетофон (DIN конектор), црно-бели видео-сигнал са PAL синхронизацијом (DIN конектор), и UHF антенски излаз (RCA конектор).

Основни циљ рачунара „галаксија“ био је да људи могу сами да га саставе (кад већ не могу да га купе) и након тога корисно употребе и да на њему играју игрице. Техничке карактеристике биле су такве да су људи могли самостално, помоћу лемилице и мало стрпљења, саставити свој први рачунар. О томе колико је већ било популарно самостално састављање рачунара говори и прилог из 130. броја часописа *Галаксија*, где је у тексту „Клуб у Новом Саду“ представљен петнаестогодишњи Иван Зиндовић, који је на трибини организације Центра за изучавање друштвеног и технолошког прогреса и уз присуство стотинак ученика средњих школа приказао микрорачунар сопствене израде. Успео је да састави рачунар чија је вредност била данашњих 6.000 динара.


**Слика 1.** Рачунар „галаксија“ (Извор: https://www.limundo.com/kupovina/Racunari-i-oprema/Desktop-racunari/Retro-racunari-i-oprema/Racunar-GALAKSIJA/70103817)

Све ово допринело да „галаксија“ као једини домаћи рачунар осамдесетих година прошлог века уђе у домове грађана и омогући популаризацију коришћења рачунара у кућне сврхе.

## 2. Магија часописа *Галаксија*

*Галаксија* је научнопопуларни часопис који је излазио једном месечно и имао 66 страна. Цена часописа од јануара до јуна 1983. године износила је тадашњих 45 динара, а након тога 50 динара. Главни и одговорни уредник био је Гаврило Вучковић, а уредништво су сачињавали:

– заменик главног и одговорног уредника Есад Јакуповић;
– помоћник главног и одговорног уредника Танасије Гавриловић;
– уредници Александар Милинковић и Јова Регасек.

Почео је да излази 1. марта 1972, као *Часопис за ваздухопловство, астронаутику и истраживање будућности*, издање НИП *Дуга*. Од новембра 1972. године (број 8) долази до репозиционирања часописа и у новом заглављу пише „Часопис за популаризацију науке“, а тако остаје све до краја периода који посматрамо у раду. Уколико погледамо ко је

све писао за часопис, очигледна је уредничка намера да се *Галаксија* удаљи од псеудонауке и да заузме објективан став према одређеним темама. О томе сведочи и издавачки савет часописа који су сачињавали, поред осталих, и шест доктора наука, као и два професора Универзитета.

У том периоду актуелне теме у медијима биле су мисија Аполо, нуклеарни погони, рачунари и сл. Хладни рат се водио науком и то је изазвало велико интересовање популације за теме из ових области. Максимални тираж који је достигла *Галаксија* био је око 85.000 примерака; почело је од 50.000 примерака (Planeta 2021) мада је било тренутака када је тираж падао на 30.000–35.000 примерака.

Концепт часописа био је такав да је увек постојала тема броја која је представљена на насловној страни и детаљно обрађена у самом часопису кроз неколико текстова. Распоред осталих тема разликовао се од броја до броја. Константно је била присутна секција „Игре, хоби, уради сам“, која се увек налазила на крају и најчешће заузимала шест страница часописа.

Часопис *Галаксија* био је један од омиљених о чему сведоче и писма читалаца, али и сам начин на који је уредништво непосредно комуницирало са њима. Отвореност је била толика да су уредници чак препоручивали стране часописе који су покривали исту тематику, без обзира на то што су конкуренти. Директан доказ налазимо у броју 132 у тексту „Страни часописи“, где стоји: „Препоручили бисмо вам два: *MC microcomputer* и *Micro and personal computer*. Претплата износи око 30.000 TRL лира, а адресе на које треба писати су...“

## 3. Методолошка замисао и инструменти истраживања

Када је потребно урадити нешто први пут, када се уводи новина било у веће друштво или на нивоу неке мање заједнице – промена започиње комуникацијом: како представљена новина, каква се порука шаље и на који начин – то су важна питања која захтевају одговор. Будући да је увођење рачунара у свакодневни живот становника Југославије био један од веома важних задатака за даљи развој индустрије, економије, тиме и целе заједнице, од изузетног је значаја анализа комуникације која је претходила великом успеху првог домаћег рачунара намењеног широкој употреби, а који су корисници могли сами саставити. Важно је разумети поруке, њихов интензитет и фреквенцију, њихов квантитет и квалитет – посебно уколико желимо да у будућности

поновимо успех увођења новина попут ове. Под квалитетом писања као наводи Стјепан Гредељ у књизи *С ону страну огледала* (Гредељ 1986) о изабраним проблемима подразумева се: број написа обухваћених узорком, њихова величина и комуникацијска вредност. Прва два показатеља су непосредно уочљива и лако мерљива. Комуникацијска вредност је изведени показатељ који је резултат процене значаја који се придаје тексту с обзиром на његов положај у структури листа. Најважнијим се процењују текстови који заузимају ударне странице листа, насловну, прву и другу, док су текстови на осталим страницама мање важни.

Под анализом садржаја подразумева се истраживачки поступак којим се настоји изградити системска искуствена евиденција о друштвеној комуникацији (Manić 2017). Анализа садржаја највише се бави анализом секундарне грађе – оне која није настала за потребе истраживања већ потпуно независно од ње. Изабрана је као метод јер омогућава и квантитативну и квалитативну анализу – у овом случају анализу садржаја часописа *Галаксија*, тада најзначајнијег научнопопуларног часописа на југословенском тржишту. Један је од ретких који су у то време системски покривали теме из области развоја рачунара и технологије. Успех који је постигао часопис овом акцијом оправдава избор овог часописа за један од извора грађе. Као извори грађе за потребе проучавања садржаја часописа о истоименом југословенском рачунару коришћени су:

- постојећа историјска и научнотеоријска литература о југословенском друштву током 1983. и 1984. године;
- искази и сведочења људи који су били учесници изградње рачунара „Галаксија“ и писали текстове за часопис *Галаксија* (Воја Антонић, Дејан Ристановић, Станко Поповић);
- методолошка литература о анализи садржаја;
- изабрана издања и текстови часописа *Галаксија* на којима се примењује анализа садржаја.

Сви обрађени примерци часописа пронађени су и купљени преко платформе Купиндо као колекционарски примерци, с обзиром на то да до дигиталних верзија издања није било могуће доћи. Како су примерци набављени у прилично лошем стању: листови су били влажни и изложени дугорочном дејству дима, а узевши у обзир тип повеза, било какво растављање би потпуно уништило примерке часописа. Услед недостатка професионале опреме и знања у области рестаурације

папира, одлучила сам се за примену оригиналне методе анализе садржаја која се састоји од читања текстова и ручном пребројавања термина и одабраних фраза.

Узорком за анализу садржаја обухваћени су сви бројеви часописа *Галаксија* који су изашли током 1983. године од броја 129 (јануар 1983. године) до прва два броја 1984. године: 141 и 142 (изузев броја 136, који није било могуће пронаћи на тржишту, а приступ библиотекама је отежан услед рестрикција изазваних глобалном пандемијом ковида). Временски оквир дефинисан је у складу са предметом истраживања, односно појавом рачунара „галаксија“, који представља фокус истраживања. Како је првобитни план за излазак рачунара био последњи квартал 1983. године, циљ је био да се испрати начин увођења теме у циљну групу, интензитет писања пре покретања акције, промена интензитета писања током акције, конотације која се даје теми, као и утицаја писања на резултате саме продајне акције. Стога је, као идеалан, изабран период од јануара 1983. године закључно са фебруаром 1984. године.

По речима Дејана Ристановића, пројекат „Галаксија“ за њега је почео 24. јуна 1983. године, када му се јавио Зоран Васиљевић са идејом да погледа мали кућни рачунар који би читаоцима могао да буде интересантан. Због тога период пре јунског издања часописа *Галаксија* суштински можемо посматрати на један начин, као креирање текстова које су аутори органски наменили одређеној теми. Са друге стране, период након јуна, када се часопис укључио у заједнички пројекат на другачији начин – више је усмерен на акцију и пропагирање одређених идеја са унапред дефинисаним продајним циљем. Дејан Ристановић на свом блогу говори о томе: „… питао сам се колико ће часопис бити расположен да се ‘уплете’ у нешто што му није основна делатност, али се показало да је интересовање огромно: Јован Регасек, тада уредник блока у коме сам писао о рачунарима, и Гаврило Вучковић, дугогодишњи главни и одговорни уредник *Галаксије*, одмах су уочили значај ове идеје и показали велику спремност да се из све снаге ангажују на реализацији. Уз једну важну измену: пројекат неће бити објављен у *Галаксији*, већ у специјалном издању „Рачунари у вашој кући“, које је требало да изађе крајем 1983. године“ (Ristanović).

У анализи су коришћене две основне јединице: теме и симболи. Анализом тема утврђено је који су најважнији предмети написа који доминирају у изабраном периоду и у каквом су они односу према појавама и догађајима којима су посвећени; затим да ли нешто претходи

акцији посвећеној склапању рачунара, дешава се паралелно са њом или након ње. Овај поступак захтева интерпретацију грађе, а створене процене се проверавају у другим изворима података – пре свега, у научној литератури везаној за тај период и ту тему, као и у сведочењима самих учесника догађаја.

Анализом симбола покушало се утврдити који су појмови у оптицају на страницама часописа *Галаксија* током 1983. и почетком 1984. године. У вредносно-мотивационом деловању средстава комуникације поједине речи су саме по себи носиоци одређене поруке, па и читаве теме. Мерена је учесталост ових појмова и њихова усмереност (да ли је позитивна, негативна или неутрална). Такође, стављање речи у одређен контекст од суштинског је значаја када је у питању утицај на промене у одређеним процесима – у овом случају наручивање и самостално склапање рачунара „галаксија“.

Пре него што је урађен категоријални систем за анализу, извршено је прелиминарно истраживање на мањем узорку текстова из посматраног периода, након чега и груписање категорија по логичкој повезаности. Сусрећемо се са двема категоријама:

1. категорије садржаја – с обзиром на то шта је речено;
2. категорије облика излагања садржаја – с обзиром на то како је нешто речено.

Категоријама садржаја испитују се предмет и смер порука. Предмет одговара на питање о чему говори порука. Смер се односи на симболе и утврђује да ли су они позитивно наклоњени према теми, неутрални или негативни. У истраживању су разликовани текстови у оквиру секције „Игре, хоби, уради сам“, која се увек налазила на зачељу месечника (последњих шест страна), текстови теме која је представљена и на насловној страни часописа, као и текстови који се налазе у издању у вези са темом, а нису везани ни за секцију ни за тему броја. Под квантитетом писања о изабраним проблемима подразумевају се број написа обухваћених узорком, њихова величина и комуникацијска вредност. Изабране теме су следеће:

– човек и машина, робот – као једна тема због илустративног спајања човека и машине у робота;
– вештачка интелигенција – као тема о којој се озбиљно расправљало и у свету и код нас и која је побуђивала различите контроверзе;
– џепни, кућни рачунари, компјутери као једна тема – јер се често користе као синоними;

– микрорачунари као одвојен сегмент због другачије употребе термина, пре свега у контексту пословне и професионалне употребе рачунара.

Што се тиче пошиљалаца, оних којима је намењена порука и самог дејства поруке, извучени су посредни подаци, проверени и допуњени подацима пронађеним у извештајима и анализама доступним из тог периода. Пошиљаоци поруке су, пре свега, аутори текстова и уредништво часописа *Галаксија*. Примаоци поруке су читаоци овог часописа заинтересовани за рачунаре и иновације на тржишту и нису желели да буду одвојени од светских трендова.

Одговор на питање како крије се у часопису *Галаксија*, у то време једном од најпопуларнијих часописа који је обрађивао теме из области технике и рачунара. Писањем текстова који су намењени разбијању митова о самим рачунарима, промовисању самоградње рачунара, као и секцији „Уради сам“, активно је промовисана тема.

## 4. Квантитет писања о изабраним проблемима

Број написа обухваћених узорком и њихов обим непосредно су уочљиви као број и површина текстова, а комуникацијска вредност је изведени показатељ, резултат процене значаја који се придаје тексту с обзиром на његов положај у структури листа. Највећи пондер додељен је тексту који је промовисан на насловној страни, а који је тема броја и у оквиру теме му је посвећено највише страна. Текстови на осталим страницама мање су важни. Укупно има пет пондера приказаних у табели 1. Уколико једна од одабраних тема („човек и машина, робот“, „вештачка интелигенција“, „џепни, кућни рачунари, компјутери“, „микрорачунари“) припада некој од категорија приписује јој се одређени пондер.

Узорком је обухваћено 13 бројева. Сваки број *Галаксије* имао је 66 страна. Број страна посвећен секцији „Игре, хоби и уради сам“ у просеку је шест (креће се између пет и седам) – дакле, 9% часописа посвећено је темама од интереса за овај рад. Поред ове секције, текстови намењени темама рачунара, робота и вештачке интелигенције појављују се у различитим сегментима у још седам бројева, што чини 53% обрађених издања. У тих седам бројева је укупно 12 текстова посвећено изабраним темама: човек и машина, робот, вештачка интелигенција, џепни, кућни рачунари, компјутери и микрорачунари. Заступљеност излажемо у табели 2.

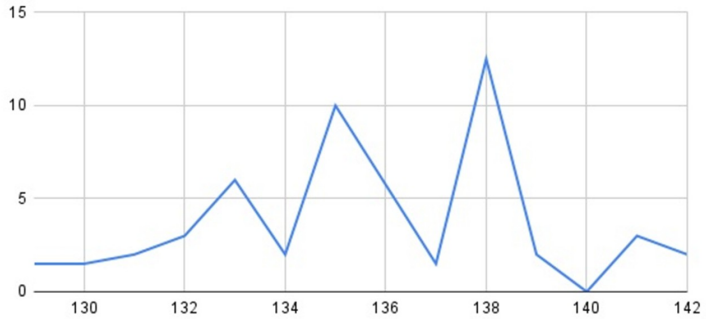
| Секција | пондер |
|---|---|
| Тема броја | 5 |
| Друга страница | 4 |
| Трећа страница | 3 |
| Четврта страница | 2 |
| Остале странице | 1 |

**Табела 1.** Пондерисање текстова

| Број Галак. | Бр.стр. за игре, хоби | пондер | Бр.стр. за склап. рачунара | пондер | Додатни текстови посв. теми | Бр.стр. за дод. текстове | пондер | Ук.бр. страна посв. теми |
|---|---|---|---|---|---|---|---|---|
| 129 | 5 | 5 | 1,5 | 1,5 | 0 | 0 | 0 | 1,5 |
| 130 | 5 | 5 | 1,5 | 1,5 | 0 | 0 | 0 | 1,5 |
| 131 | 6 | 6 | 2 | 2 | 1 | | | 2 |
| 132 | 6 | 6 | 1,5 | 1,5 | 1 | 1,5 | 1,5 | 3 |
| 133 | 7 | 7 | 2 | 2 | 2 | 4 | 4 | 6 |
| 134 | 6 | 6 | 0,5 | 0,5 | 1 | 1,5 | 1,5 | 2 |
| 135 | 6 | 6 | 1 | 1 | 2 | 9 | 45 | 10 |
| 137 | 6 | 6 | 1,5 | 1,5 | 0 | 0 | 0 | 1,5 |
| 138 | 6 | 6 | 2,5 | 2,5 | 4 | 10 | 50 | 12,5 |
| 139 | 6 | 6 | 2 | 2 | 0 | 0 | 0 | 2 |
| 140 | 7 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 141 | 6 | 6 | 2 | 2 | 1 | 1 | 1 | 3 |
| 142 | 6 | 6 | 2 | 2 | 0 | 0 | 0 | 2 |

**Табела 2.** Пондери за странице посвећене теми

Уочавамо да нема броја у коме није обрађена барем једна од тема: човек и машина, робот, вештачка интелигенција, џепни, кућни рачунари, компјутери и микрорачунари. На графикону (слика 2) уочавамо неравномерну расподелу укупног броја страна посвећених темама и јасан тренд пада почетком 1984. године, након изласка специјалног издања намењеног склапању рачунара. Врхунац се достиже у октобарском броју 138, где је тема броја била рачунар, па је и очекивано да је максимална пажња посвећена изабраној теми. Такође, на графикону се јасно види заинтересованост уредништва за тему и пораст заступљености и пре сазнања о постојању рачунара „галаксија“ и плана акције.




**Слика 2.** Укупан број страна посвећених темама по бројевима *Галаксије*.

Што се тиче насловне стране, укупно постоје три броја (23%) на којима је присутан један од ова три елемента (насловница садржи реч рачунар, насловница садржи фразу *вештачка интелигенција* и насловница садржи графички приказ рачунара). У броју 138 тема је рачунар, те је ова реч присутна и у самом наслову, као и у графичкој илустрацији на насловној страни. Конотација тема никада није негативна: или је позитивна или неутрална. У табели 3 приказана је процентуална заступљеност тема на насловној страни *Галаксије*.

| Опис | Проценат |
|---|---|
| Насловница садржи реч *рачунар* | 7,69% |
| Насловница садржи фразу *вештачка интелигенција* | 7,69% |
| Насловница садржи графички приказ рачунара | 23,08% |

**Табела 3.** Процентуална заступљеност тема на насловној страни *Галаксије*.

У истраживању је обрађено укупно 47 текстова. Сваки часопис је у складу са методом анализе садржаја прочитан, пребројане су све обрађене фразе и речи и подаци су складиштени у табеле. Поред тога обрађена је и величина текстова што је специфичан показатељ публицитета који се даје одређеној проблематици и теми у одређеном времену (табела 4). За *Галаксију* као научнопопуларни часопис карактеристично је да није био намењен за једнократну употребу (за

разлику од дневних издања новина), већ су се примерци чували, изнова читали и у одређеним сегментима служили као уџбеници и приручници. Тиме је вредност сваког текста већа јер је утицај далекосежнији, не представља само тренутну забаву и разоноду. Заступљеност теме о склапању рачунара је константна. Посвећена јој је у просеку једна и по страна у сваком броју. Максималан број страна посвећен теми је 2,5, док само један број часописа није имао ниједан текст .

| бр. Галаксије | човек и машина, робот | ВИ, интелигентне машине | кућни рачунари, компјутери и сл. | микро рачунар | демократизација компјутера |
|---|---|---|---|---|---|
| 129 | 0 | 1 | 16 | 5 | 0 |
| 130 | 0 | 0 | 13 | 4 | 0 |
| 131 | 0 | 4 | 40 | 2 | 1 |
| 132 | 19 | 1 | 34 | 2 | 0 |
| 133 | 0 | 2 | 133 | 1 | 0 |
| 134 | 0 | 0 | 30 | 0 | 0 |
| 135 | 184 | 8 | 60 | 1 | 0 |
| 137 | 1 | 1 | 43 | 1 | 0 |
| 138 | 5 | 14 | 129 | 32 | 0 |
| 139 | 0 | 0 | 30 | 0 | 0 |
| 140 | 0 | 0 | 25 | 0 | 0 |
| 141 | 5 | 0 | 53 | 11 | 0 |
| 142 | 0 | 0 | 53 | 2 | 0 |
| збир | 214 | 31 | 659 | 61 | 1 |
| просек | 16,46 | 2,38 | 50,69 | 4,69 | 0,08 |

**Табела 4.** Заступљеност теме у обрађеним издањима часописа *Галаксија*.

У оквиру табеле 4 уочава се стабилан тренд и присутност текстова у вези са склапањем рачунара. Са друге стране, присутност у оквиру текстова ван тог одељка је веома нестабилна, те је јасно да је део системских настојања уредништва да се скрене пажња на одабрану тему. Приметно је да се у првим месецима 1983. године изабране теме нису појављивале изван дела који се тиче хобија, али од лета се интензитет умногоме мења, као што је приказано на графикону (слика 3).

Кључне речи и фразе које се најчешће јављају у оквиру анализираних текстова:

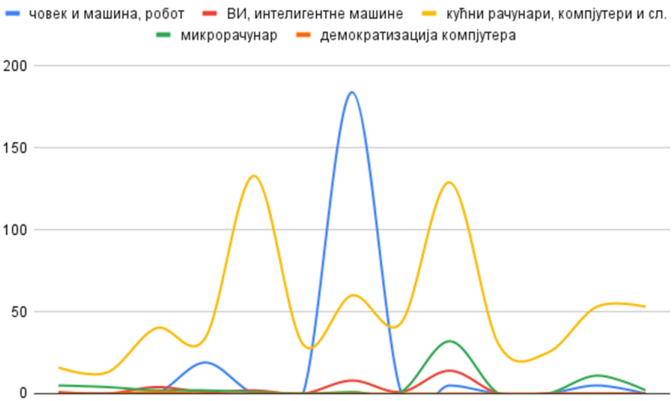
1. Џепни рачунар, кућни рачунар и компјутер (користе се као синоними) – укупно 659 пута, у просеку 50 пута по броју.
2. Човек и машина, робот – укупно 214 пута, у просеку 16 пута по броју.
3. Микрорачунар – 61 пут, у просеку четири пута по броју. Термин микрорачунар се више користи за пословну примену рачунара и није одомаћен као израз када се разговара о кућној и личној (персоналној) употреби рачунара – стога је у анализи и издвојен као кључна реч.
4. Вештачка интелигенција – 31 пут, у просеку 2,38 пута по броју.
5. Демократизација рачунара – свега једанпут, у броју 131, када су и најављени клубови љубитеља рачунара.

Пре тога, у броју 130 уводи се појам клубова као одличан начин организовања љубитеља рачунара. Узевши у обзир време када је лист настао, реч *демократизација* није била популарна и, без обзира на њену потпуно бенигну употребу у овом контексту, уредништво је никада касније није употребило.

Кроз саму анализу кључних речи очигледно је наглашена лична/персонална употреба рачунара и бирани су термини и фразе блиски читаоцима, речи које не стварају одбојност. Такође, дата је предност односу човека и машине, као и роботима, а не самој вештачкој интелигенцији иако се тада та грана науке веома интензивно развијала. Ову тврдњу илуструје и чињеница да у броју 129 пише: „Централни процесор је срце и мозак рачунара, део машине који мисли.“ На графикону (слика 3) приказана је расподела кључних речи по бројевима и приметан је утицај теме броја на заступљеност и учесталост кључних речи. Уочава се и да су од лета до пред излазак броја специјалног издања намењеног састављању рачунара „галаксија“ интензивно покривене кључним речима, што указује на то да је уредништво припремало терен за акцију.

## 5. Небо је граница: најсмелији сан – 1.000 наруџбеница; пристигло 8.000

У СФРЈ је 1983. и 1984. године био регулисан увоз робе чија вредност износи тадашњих 1.500 динара и више. Како Танјуг јавља 22. VII те године, у првој половини године више је него преполовљен југословенски дефицит у трговини са западом (са 2,12 на 0,99 милијарди долара) захваљујући смањеном увозу и повећаном извозу. Поред тога, Уставни суд је приговорио Савезној скупштини због закона којим се путовање




**Слика 3.** Расподела кључних речи и фраза.

у иностранство условљава депозитом од 5.000 динара (уведено октобра 1982), чиме су и честа путовања у иностранство била тешко изводљива. Због царинских мера увоз комплетних рачунара био је немогућ. Југославија је 1983. године службено банкротирала иако то никад својим грађанима није објавила (но објавили су други), те је престала да плаћа све обавезе према иностранству, што је резултирало великим несташицама. Са друге стране, домаћа индустрија није успевала да довољно брзо и по прихватљивим ценама развије рачунаре за кућну употребу. Интересената је било, персонални рачунари су били „врућа тема“ међу љубитељима технике који су се довијали на свакакве начине да дођу барем до компонената како би склопили рачунар и отпочели програмирање или уживали у игрицама. Уредништво листа, а пре свега три човека најзаслужнија за праћење теме – Станко Поповић, Дејан Ристановић и Воја Антонић, сваки кроз своју мрежу контаката, своју делатност, разговоре са читаоцима *Галаксије*, кроз праћење локалних дешавања, кроз питања која су им стизала, уочили су пораст интересовања за рачунаре и одлучили да реагују на повећању тражњу.

Приликом читања текстова запажа се веома отворена комуникација са читаоцима по принципу питања и одговора који се објављују у оквиру одељка намењеног за то. Примећује се да уредништво прати и

локална дешавања по градовима, посебно клубове љубитеља рачунара, као и да радо посећује њихове догађаје и објављује вести о њима. То омогућава грађење мреже контаката и стално примање информација из свих крајева земље – непрекидно опипавање пулса читалаца и дешавања. Управо на овај начин су уредници и новинари сазнали које теме занимају читаоце и осмислили како да их представе. Поред тога, кроз садржај часописа *Галаксија* веома јасно се види да су новинари пратили сва страна издања, преводили чланке и прилагођавали их за домаће тржиште. У интервјуу који је дао за Југопапир, Дејан Ристановић каже следеће о страним часописима у вези са рачунарима који су се могли набавити у Србији: „Прави квалитет који нам нуде те стране ревије изванредни су прикази нових типова рачунара и периферијске опреме, али то је ипак премало за цену листа и поштанских трошкова, који се плаћају у девизама“ (Yugopapir 2014).

Према подацима Републичког завода за статистику, просечна месечна плата у Србији тада је износила 15.161 динара (Republički zavod za statistiku 2006), а специјално издање *Галаксије* које се бавило склапањем рачунара коштало је 200 динара. То значи да је за тај посебан број *Галаксије* требало издвојити око 1,3% просечне плате. Године 1983. цена килограма јабука износила је 35 динара, а килограм чоколаде коштао је 257 динара (Milićević 2018) – стога можемо сматрати, на основу података Макроекономије, да су се за цену *Галаксије* могли покрити дневни трошкови просечне породице у Србији.

Како и сам Ристановић истиче, „то је имало своју сврху у тренутку када је рачунаре требало популарисати и када у земљи није било довољно програма...“ (Yugopapir 2014). На почетку најважнија је била спознаја да је самоградња могућа, а то је оно што је претходило писању *Галаксије*. Сва интегрисана кола могла су легално да се увезу поштом из иностранства, а набавка штампаних плоча, тастатура и кутија могла се организовати у Југославији. План је био да програмирају EPROM меморије, јер су пошли од претпоставке да ће стићи максимално 100 захтева, најхрабрији су очекивали 1.000 – међутим, стигло је 8.000 захтева.

По речима самог Ристановића: „Да би оваква акција успела, треба је неколико месеци припремати и пратити кроз текстове у *Галаксији*, па је тако септембра 1983. објављен мој текст у коме је самоградња представљена. Текст је пратила прелиминарна наруџбеница, а одзив је био невероватан: преко хиљаду читалаца *Галаксије* изражава жељу да сагради свој први компјутер! (...) Када нека акција добро крене,

људи који у њој учествују имају веома јак мотив да се и даље труде: Воја Антонић је следећих месеци безброј пута побољшао хардвер и софтвер рачунара, а *Галаксија* је покушавала да пронађе најпогоднији начин да организује набавку комплета за самоградњу, ослањајући се на „малу привреду“, која је тих година тек настајала. Тако смо уговорили да ‘Мипро’ и ‘Електроника’ из Буја, у сарадњи са Институтом за електронику и вакуумску технику, испоручују штампане плоче, тастатуре и маске, да ‘Микротехника’ из Граца шаље чипове, а да *Галаксија* прикупља наруџбенице и програмира EPROM.“

То је оно што је *Галаксија* и урадила. Одлично је пратила тему и креирала потражњу изабравши сјајне ауторе – врсне познаваоце теме, блиске трендовима који су постојали у свету, а који пишу на начин који је близак читалачкој публици. Исплатило се инсистирање на самоградњи, „уради сам“ и активном учешћу у састављању рачунара, које се годинама протезало кроз *Галаксију*, а читаоци су били уверени да могу сами узети лемилицу у руке и саставити сопствени рачунар и почети нешто корисно да раде на њему. Одлична упутства, сликовити прикази, позитивна конотација рачунара били су ветар у једра целом пројекту. Ево неколико примера:

– У 131. броју, у тексту под називом „Комуникација са рачунаром“, постоји шест илустрација процеса комуникације са рачунаром. Као што се може видети и на слици 4, на свакој илустрацији је рачунар насмејан и расположен за комуникацију, и сарадња са човеком представљена је као тимски рад.
– У броју 135, у тексту са маштовитим насловом „Полетео вараждински ‘галеб’“ најављује се први југословенски микрорачунар намењен личној и кућној употреби – односно његов улазак у серијску производњу. Наводи се и цена рачунара од 89.035 динара и констатује како је она доста висока за оно што се нуди. Уз текст иде и фотографија рачунара уз објашњење: „По спољашњем изгледу, ‘галеб’ се може мерити са многим рачунарима стране производње.“
– У 139. броју, под насловом „Све моћнији“, приказана је и графика рачунара „галаксија“. Прва порука исписана на екрану гласила је: „И ти можеш да направиш рачунар *ГАЛАКСИЈА*“ – боља мотивација од те није потребна.

Проблеми настају у реализацији самог пројекта – логистика и набавка су прве отказале послушност. „Прво није било витропласта, па није било тастера, па није било довољно новца за производњу (плаћало


**Слика 4.** Комуникација са рачунаром.

се поузећем), па није било чипова... Марфијеви закони су се и те како показали на делу: цене чипова су годинама падале да би, почетком 1984, нагло скочиле. Статички RAM-ови су баш тада поскупели, појавили су се проблеми у набавци EPROM-а од 4 килобајта, Z-80 је појефтинио, али је Галаксијин Z-80A поскупео. Програмирање EPROM-а је било мало успорено, „Микротехника“ је каснила неколико месеци, MIPRO и „Електроника“ готово читаву годину, а сви ти проблеми су се сурвали на редакцијски телефон *Галаксије*. Хиљаде телефонских позива су основни узрок како за закашњење „рачунара 2“, тако и за успорен рад на даљем развоју рачунара „галаксија“ (Ristanović).

Једини начин да се посао доведе до краја подразумевао је много директног рада и контаката са наручиоцима, као и много стрпљења

како би се решили проблеми у вези са набавком и логистиком. Почетком 1984. године у помоћ је притекао и радио, те су снаге удружили Зоран Модли, Дејан и Воја. Сваког петка у етру је била нова епизода, а теме су биле популарне и занимљиве ширем кругу слушалаца – углавном игре и демонстрације: *Замак*, *Jumping Jack*, *Мастермајнд*, *Еволуција*, *Галактички рат*, *Хамураби*, *Биоритам*.

Уредништво и творци рачунара „галаксија“ нису били задовољни софтвером који су направили крајњи корисници. Од више хиљада наруџбеница *Галаксије*, стигло је свега пет-шест програмских решења вредних објављивања. Закључак редакције био је да је креирање програма знатно захтевнији посао од састављања рачунара, те да је битно да се креаторима омогући и комерцијална надокнада за њихов рад и труд.

## 6. Закључак

Данас је тешко замислити ситуацију у којој не можете да одете у продавницу и купите оно што вам треба: било да је реч о рачунару, мобилном телефону или неком софтверском решењу. Не морате чак ни да идете – све је доступно „на клик“ и свака куповина може да се обави преко интернета, а да већ следећег тренутка сачекате испоруку на кућну адресу. Приказани случај првог домаћег рачунара „галаксија“ за кућну употребу показује да је и у временима великих ограничења и рестрикција могуће наћи решење уколико се упосле знање и умеће, тј. ако се креативно приступи решавању проблема. Улога која је преузета у том случају је активна, а не пасивна – конзумерска. Човек је у овом случају активан стваралац, а јасно се показује да медији (у овом случају научнопопуларни часопис *Галаксија*) могу да подстакну људе на конкретну акцију и улију им довољно вере и инспирације за активно деловање. Циљ писања било је покретање људи на акцију, без почетне жеље за комерцијалним успехом – који је, на крају, ипак остварен. У интервјуу који је дао за портал „Стартит“, одговарајући на питање због чега је *Галаксију* бесплатно уступио људима, Воја Антонић објашњава: „Због чега сам и све друго радио. Мени је то био циљ. Касније, кад је објављен први број часописа, појавила се нека фирма са Истре која је дистрибуирала компоненте, увозила их из Аустрије и продавала заједно са програмираним EPROM-има. Они су финансијски лепо прошли. Када сам на крају израчунао колико су тога они продали, испоставило се да су зарадили више од милион тадашњих марака. То би, теоретски,

требало да ме потресе, али није ме потресло. Њима је то био циљ, мени је недостајало похлепе да бих тако нешто урадио“ (Startit, 2017).

Сам однос са читаоцима *Галаксије*, који се одвија кроз дијалог (у то време асинхрон јер је захтевао дописивање путем дописница или телеграма), указује на постојање дискурса о састављању рачунара, одабира најбољих опција у одређеном ценовном рангу, истраживања потенцијалних решења и подстицање стваралачког чина. Такав вид комуникације са уредништвима часописа данас је готово потпуно нестао. Како се коментари остављају путем интернета и различитих платформи, њихово творци не употребе довољно времена и енергије да артикулишу своје ставове и мисли, и остваре конструктиван дијалог. Са друге стране, уредништво ни издалека не посвећује времена и енергије ставовима читалаца, њиховим питањима. Модерација коментара одвија се на потпуно другачијем нивоу и у оквиру других тимова, најчешће оних који су задужени за друштвене мреже. Питање објављивања и необјављивања коментара и утисака читалаца исто је као што је било и 1983. године – нажалост, никада нећемо сазнати колико тога је остало необјављено и зашто.

На основу сведочења актера, као и података истраживања, јасно је да рачунар „галаксија“ никада не би постигао успех без системске подршке часописа *Галаксија*. Пројекат је добар пример како приступити решавању проблема на креативан и довитљив начин у тренуцима када су постојала објективна ограничења која је наметнула држава.

## Литература

Le Parisien. 2000. *Dictionnaire Sensagent.* Retrieved from Dictionnaire Sensagent. http : / / dictionnaire . sensagent . leparisien . fr / List % 20of % 20computer % 20systems % 20from % 20the % 20Socialist % 20Federal % 20Republic%20of%20Yugoslavia/en-en/.

Manić, Ž. 2017. *Analiza sadržaja u sociologiji.* Beograd: Čigoja štampa.

Milićević, D. 2018. *Makroekonomija.* Accessed on March 18, 2018. Retrieved from Makroekonomija. http://www.makroekonomija.org/0-dragovan-milicevic/sfrj-%E2%80%93-iluzija-ili-stvarnost-boljeg-zivota/.

Planeta. 2021. *Planeta.rs.* Accessed on April 04, 2021. Retrieved from Planeta.rs. http://www.planeta.rs/54/04_temabroja1.htm.

Republički zavod za statistiku. 2006. *Zarade u Republici Srbiji 1965–2005.* Technical report. Retrieved from Republički zavod za statistiku. https://pod2.stat.gov.rs/ObjavljenePublikacije/G2006/Pdf/G20066004.pdf.

Ristanović, D. *Dejan Ristanovic.* N.d. Retrieved from dejanristanovic.com. https://www.dejanristanovic.com/galaks.htm.

Startit. 2017. *Startit.* Accessed on July 15, 2017. Retrieved from Startit.rs. https://startit.rs/voja-antonic-danas-su-kompjuteri-bela-tehnika-a-nekada-su-bili-cudo-prirode/.

Yugopapir. 2014. *Yugopapir.com.* Accessed in August 2014. Retrieved from Yugopapir. http://www.yugopapir.com/2014/08/dejan-ristanovic-intervju-sa-mladim.html.

Гредељ, Стјепан. 1986. *С ону страну огледала. Истраживање промена модела комуникације у југословенском друштву на основу анализе садржаја писања листова Борбе и Политике у периоду од 1945. до 1975. год.* 232 стр. Београд: Истраживачко-издавачки центар ССО Србије.

# Увод у Дигиталну хуманистику: радионице за имплементацију ‘удаљеног читања’ у истраживачкој пракси, Тршић – 4–8. 12. 2023.

Милица Рабреновић
milica.rabrenovic@isj.sanu.ac.rs
ORCID: 0000-0002-4441-4884
*Институт за српски језик САНУ*
*Београд, Србија*

**РАД ПРИМЉЕН:** 17. април 2024.
**РАД ПРИХВАЋЕН:** 29. мај 2024.

## 1. Увод

Универзитетска библиотека „Светозар Марковић” и Друштво за језичке ресурсе и технологије – JeРТех организовали су први семинар „Увод у Дигиталну хуманистику: радионице за имплементацију ‘удаљеног читања‘ у истраживачкој пракси”, који је реализован у Тршићу од 4. 12. до 8. 12. 2023. године. Семинар је спроведен кроз сарадњу његових аутора и Министарства науке, технолошког развоја и иновација Републике Србије. Полазници семинара били су студенти основних, мастерских и докторских студија. Циљ овог семинара био је да се студенти филолошких и хуманистичких наука упознају с методама дигиталне хуманистике, припреме и обраде текста за примену техника ‘удаљеног читања’. Семинар је обухватао део с предавањима и практичан део, односно радионице.

Првог дана семинара одржана су два предавања, потом је у поподневним сатима уследила примена теоријског дела. Прво предавање „Увод у дигиталну хуманистику” одржао је др Василије Милновић. На почетку излагања, др Милновић се осврнуо на резултате њихових досадашњих пројеката, упознао је полазнике с пројектом у оквиру COST Action CA16204 *Distant Reading for European Literary History* (2018–2022) и истакао важност дигиталне хуманистике. Говорио је о значају развијања језичких технологија, с посебним освртом на тзв. "мале језике" (нпр. Исланд, Естонија или Израел где се улажу значајна средства у развоју тог процеса) и њиховој потенцијалној примени у

лексикографији, настави српског језика и књижевности, превођењу и другим областима.

Предавање проф. др Цветане Крстев било је посвећено корпусу *SrpELTeC*. Проф. др Крстев упознала је полазнике с процесом настајања овог корпуса и критеријумима које је било неопходно испунити да би српски роман у датом периоду постао део корпуса (равномерна заступљеност и мушких и женских аутора; равномерно покривање изабраног периода (1840–1920); разноврсност по питању дужине; позната и призната дела, али и непозната; разноврсност аутора). Било је речи о изазовима с којима су се сусретали током формирања српске потколекције, као и о процесу дигитализације првих издања (или најстаријих доступних). Процес дигитализације чине следећи кораци: скенирање, оптичко препознавање карактера, корекција, основна анотација, уношење метаподатака, аутоматска напредна анотација.

Полазници семинара упознати су на који начин се врши корекција и обележавање текста, као и како се уносе метаподаци, те су први дан радионице добили текст на којем су радили корекцију пратећи упутства за обележавање, а током осталих дана уносили су метаподатке и радили су на анотацији језичких ентитета у том тексту.

Другог дана је др Александра Тртовац, у свом предавању „Универзитетска библиотека ‘Светозар Марковић’ и удаљено читање”, говорила о повезаности библиотекарства и дигиталне хуманистике, као и о њиховој улози у COST акцији. Друго предавање др Тртовац посветила је пројекту Транскрибус и дигитализацији путем мобилног телефона. Полазници су тог дана на радионици направили налоге на Транскрибусу,[1] а потом су упознати и са апликацијом *DocScan* помоћу које је могуће учитати рукописне радове и добити транскрипт текста. Веома корисно било је упознавање с напредним претраживањем у Кобису.[2] Поред изборног претраживања, фокус је био усмерен и ка командном претраживању. Једна од корисних информација била је и да се у одељку *UNPAYWALL* налазе научни чланци са отвореним приступом, као и да сва имена и презимена једног аутора (корисно уколико имамо псеудоним, а желимо да излистамо сва дела једног аутора или ако је презиме ауторке мењано услед брачних разлога) можемо наћи у делу *CONOR.SR*.

1. readcoop, приступљено 28. 6. 2024. године
2. COBISS.SR, приступљено 28. 6. 2024.

Трећег дана проф. др Крстев упознала је полазнике с појмом именовани ентитети и начином како се они обележавају у тексту. Предавање проф. др Ранке Станковић надовезало се на претходно, било је речи о обради дигиталних текстова. Учесници радионице су радили на аутоматској анотацији делова текста именованим ентитетима (места, особе, организације, професије...), дате су и одређене напомене о напреднијим врстама анотације, потом је уследила контрола и евалуација аутоматски евалуираних података.

Истог дана на радионици учесници су упознати с википодацима о српским романима, те су уносили основне податке о тексту који су добили првог дана користећи алате *OpenRefine* и *QuickStatements*. Учесници су упознати и с претраживањем података језиком *SPARQL*.

Четврти дан започет је предавањем проф. др. Душка Витаса, говорено је о процесу настајања кулинарског корпуса. Проф. др Ранка Станковић наставила је тему о корпусима доступним на NOSKE платформи друштва ЈерТех.[3] Учесници су имали прилику да се упознају с основама *CQL* језика и како да дају упите над корпусом *srpELTeC*. Радионица је била посвећена командним претрагама, почевши од једноставнијих упита, па све до сложенијих граматичких образаца и упита који укључују етикете именованих ентитета.

Последњег дана семинара учесници су слушали предавање „Иницијатива за развој отворених NLP/NLU ресурса и алата за српски језик" Слободана Марковића. Током предавања истакнути су изазови употребе вештачке интелигенције (ВИ), као и могућности и проблеми у вези с обрадом српског језика коришћењем ВИ. Проблем постоји јер модели нису суштински прилагођавани, њихова примена није практична у пословним решењима и могућности експресије су сужене. Решење је у томе да буде што више текста за обучавање. Идеално би било да је што више тога јавно доступно, под пермисивном лиценцом и да покрива оба изговора. Циљеви иницијативе за отворене NLP ресурсе јесу бољи резултати претраге и разумевање текста.

На овом семинару стечена су свеобухватна знања о методама дигиталне хуманистике и значају ове научне области. Овај семинар омогућио је боље увиде за будућа истраживања, посебно када је у питању обележавање грађе, али и рад са текстуалним корпусима, у чему је и највећа примена наученог током овог семинара. Оно што треба истаћи као најважнији резултат ових предавања и радионица јесте упознавање с колегама и развијање потенцијалне сарадње на будућим пројектима

3. noske.jerteh.rs, приступљено 28. 6. 2024.

и истраживањима. То и јесте главни разлог зашто бих као учесница топло препоручила семинар „Увод у Дигиталну хуманистику: радионице за имплементацију ‘удаљеног читања’ у истраживачкој пракси”.

## Захвалница

Овај рад финансирало је Министарство просвете, науке и технолошког развоја Републике Србије према Уговору број 451-03-136/2025-03/ 200174, који је склопљен са Институтом за српски језик САНУ.