Међународна конференција

Јужнословенски језици у дигиталном окружењу

ЈуДиг

Књига резимеа

21 - 23. новембар 2024.
Универзитет у Београду - Филолошки факултет, Србија

ОРГАНИЗАТОРИ
Филолошки факултет Универзитета у Београду
Друштво за језичке ресурсе и технологије (JePTex)

Уредници:
Проф. др Јасмина Московљевић Поповић
Проф. др Ранка Станковић

Технички уредници:
Проф. др Ранка Станковић
Др Александра Томашевић

Дизајн корица:
Анђелка Зечевић

Издавач:
Филолошки факултет Универзитета у Београду

За издавача:
Проф. др Ива Драшкић Вићановић

Штампа:
Универзитет у Београду – Филилошки факултет

Тираж:
100 примерака

Публикација доступна на:
https://judig.jerteh.rs/2024/judig-book-of-abstracts.pdf

Веб сајт: https://judig.jerteh.rs/
E-mail: judig@jerteh.rs
Графика: https://treecloud.univ-mlv.fr/

ISBN-978-86-6153-755-4

Конференцију су подржали

INSTITUTE FOR STANDARDIZATION OF SERBIA

(бронзани спонзор)

Science Fund of the Republic of Serbia

Republic of Serbia
MINISTRY OF SCIENCE, TECHNOLOGICAL DEVELOPMENT AND INNOVATION

Office for Information Technologies and e-Government

University of Belgrade
Faculty of Mining and Geology

ENVI SOFTWARE SOLUTIONS

ПРОГРАМСКИ ОДБОР

Копредседнице:

Проф. др Јасмина Московљевић Поповић, Филолошки факултет Универзитета у Београду

Проф. др Ранка Станковић, Рударско-геолошки факултет Универзитета у Београду, и Друштво за језичке ресурсе и технологије (JePTex)

Чланови:

Проф. др Агата Савари, Универзитет Париз - Сакле, Француска

Др Александра Марковић, Институт за српски језик, САНУ, Србија

Др Ана Острошки Анић, Институт за хрватски језик, Хрватска

Доц. др Балша Стипчевић, Универзитет у Београду, Филолошки факултет, Србија

Проф. Бенедикт Перак, Свеучилиште у Ријеци, Филозофски факултет, Хрватска

Др Биљана Рујевић, Универзитет у Београду, Рударско-геолошки факултет, Србија

Др Василије Милновић, Универзитетска библиотека "Светозар Марковић", Србија

Проф. др Вера Ћервиз, Универзитет у Источном Сарајеву, Филозофски факултет, Босна и Херцеговина

Др Верџиника Барбу Митателу, Истраживачки институт за вештачку интелигенцију, НЛП група, Румунска академија, Румунија

Проф. др Владан Девеџић, Универзитет у Београду, Факултет организационих наука, Србија

Проф. др Владимир Поломац, Универзитет у Крагујевцу, Филолошко-уметнички факултет, Србија

Проф.др Гордана Павловић-Лажетић, Друштво за језичке ресурсе и технологије (JePTex), Србија

Доц. др Данило Алексић, Универзитет у Београду, Филолошки факултет, Србија

Проф. др Димитар Трајанов, Универзитет светог Ћирила и Методија - Скопље, Факултет информатичких наука и компјутерског инжењерства, Македонија

Проф. др Душанка Поповић, Универзитет Црне Горе, Филозофски факултет, Црна Гора

Проф. др Душка Кликовац, Универзитет у Београду, Филолошки факултет, Србија

Проф. др Душко Витас, Друштво за језичке ресурсе и технологије (JePTex), Србија

Др Јака Чибеј, Универзитет у Љубљани, Филозофски факултет, Словенија

Проф. др Јелена Граовац, Универзитет у Београду, Математички факултет, Србија

Проф. др Јелена Јовановић, Универзитет у Београду, Факултет организационих наука, Србија

Проф. др Јелена Марковић, Универзитет у Источном Сарајеву, Филозофски факултет, Босна и Херцеговина

Доц. др Јован Чудомировић, Универзитет у Београду, Филолошки факултет, Србија

Проф. др Кети Здравкова, Универзитет светог Ћирила и Методија - Скопље, Факултет информатичких наука и компјутерског инжењерства, Македонија

Проф. др Марко Робник Шикоња, Универзитет у Љубљани, Факултет за рачунарство и информатику, Словенија

Доц. др Милош Утвић, Универзитет у Београду, Филолошки факултет, Србија

Др Михаило Шкорић, Универзитет у Београду, Рударско-геолошки факултет , Србија

Др Небојша Васиљевић, Фондација Петља, Србија

Проф. др Невена Цековић, Универзитет у Београду, Филолошки факултет, Србија

Проф. др Нелда Коте, Политехнички универзитет у Тирани, Албанија
Проф. др Оља Перишић, Универзитет у Торину, Департман за стране језике, књижевност и модерне културе, Италија
Др Параскеви Јули, Институт за језике и обраду говора, Истраживачки центар Атина, Грчка
Проф. др Петја Осенова, Универзитет св. Климент Охридски, Софија, Бугарска
Др Рада Стијовић, Друштво за језичке ресурсе и технологије (JePTex), Србија
Доц. др Саша Марјановић, Универзитет у Београду, Филолошки факултет, Србија
Проф. др Саша Модерц, Универзитет у Београду, Филолошки факултет, Србија
Проф. др Светла Коева, Бугарска академија наука, Бугарска
Проф. др Соња Ненезић, Универзитет Црне Горе, Филолошки факултет, Црна Гора
Проф. др Цветана Крстев, Друштво за језичке ресурсе и технологије (JePTex), Србија
Др Кристина Штркаљ Деспот, Институт за хрватски језик, Хрватска
Проф. др Иван Обрадовић, Друштво за језичке ресурсе и технологије (JePTex), Србија
Др Оливера Китановић, Универзитет у Београду, Рударско-геолошки факултет, Србија

ОРГАНИЗАЦИОНИ ОДБОР

Копредседнице:

Проф. др Јасмина Московљевић Поповић, Универзитет у Београду, Филолошки факултет, Србија
Проф. др Ранка Станковић, Универзитет у Београду, Рударско-геолошки факултет и Друштво за језичке ресурсе и технологије (JePTex), Србија

Чланови:

Доц. др Јован Чудомировић, Универзитет у Београду, Филолошки факултет
Доц. др Балша Стипчевић, Универзитет у Београду, Филолошки факултет
Проф. др Невена Цековић, Универзитет у Београду, Филолошки факултет
Доц. др Милош Утвић, Универзитет у Београду, Филолошки факултет
Доц. др Милица Динић Маринковић, Универзитет у Београду, Филолошки факултет
Мср Милица Иконић Нешић, Универзитет у Београду, Филолошки факултет
Др Биљана Рујевић, Друштво за језичке ресурсе и технологије (JePTex)
Др Михаило Шкорић, Друштво за језичке ресурсе и технологије (JePTex)
Др Александра Марковић, Друштво за језичке ресурсе и технологије (JePTex)
Анђелка Зечевић, Друштво за језичке ресурсе и технологије (JePTex)
Никола Гуџић, Друштво за језичке ресурсе и технологије (JePTex)

Међународна конференција *Јужнословенски језици у дигиталном окружењу – ЈуДиг* пружа могућност бројним истраживачима из области рачунарске лингвистике, језичких технологија и повезаних дисциплина да поделе своје идеје, увиде и резултате истраживања у оквиру наведених области.

Конференција има за циљ да подржи успостављање сарадње међу истраживачима, као и да обезбеди прилику за упознавање са резултатима релевантних истраживања у свим (под)областима рачунарске лингвистике и повезаних дисциплина. Посебна пажња посвећена је језичким ресурсима и технологијама које су доступне за српски и друге јужнословенске и балканске језике. У раду Програмског одбора конференције учешће је узело четрдесет троје еминентних истраживача из дванаест земаља и двадесет две научно-истраживачке институције.

За учешће на конференцији захтев је поднело педесет пет излагача, а предложене теме тицале су се великог броја најразличитијих проблема из најшире схваћене области језичких технологија. Све предложене теме и резимеи подвргнути су процесу двоструко анонимне рецензије од стране три рецензента, члана Програмског одбора. Програмски одбор и копредседавајуће одабрали су 51 рад за излагање на конференцији. (Ко)аутори одабраних радова су 84 истраживача из 34 институције из 14 земаља - педесет пет из Србије, пет из Хрватске, по четири из Бугарске и Немачке, по два из Словеније, Аустрије, Македоније, Велике Британије, Француске и Босне и Херцеговине, и по један из Грчке, Холандије, Јужне Кореје и Италије. Конференција се одржава у хибридном моду (уживо и онлајн) и укључује четири предавања по позиву, дванаест тематских сесија и четири тематски усмерене радионице.

Проминентни предавачи по позиву одржаће предавања о важним темама и актуелним сазнањима на пољу језичких технологија, те пружити увид у најновија достигнућа и најзначајније трендове који обликују дисциплину. Они долазе из четири државе: Француске (Агата Савари), Бугарске (Светла Коева), Велике Британије (Руслан Митков) и Србије (Цветана Крстев).

Поље језичких технологија данас је не само врло популарно, већ је и веома широких граница. То се види и по разноврсности тема које ће бити изложене на конференцији. Наслови педесет једног прихваћеног резимеа могу се разврстати у седам тематских целина: ИТ-обрада јужнословенских језика (5), дигитални корпуси јужнословенских језика (11), језички ресурси за јужнословенске језике (6), језичке технологије за јужнословенске језике (8), граматика и лексикон јужнословенских језика у контексту обраде природних језика (5), вештачка интелигенција, језички модели и обрада јужнословенских језика (8) и дигитална хуманистика (8).

Организациони одбор посебно се усмерио на дељење и промовисање постојећих ресурса и технологија са другим истраживачима и потенцијалним истраживачима. Током учешћа у раду четири радионице полазници ће имати прилику да стекну знања о најновијим приступима, ресурсима и алаткама из различитих области обраде природних језика, да се упознају са примерима добре праксе кроз одабране студије случаја, као и да се сретну и успоставе контакт са другим истраживачима и будућим сарадницима.

Организатори се посебно захваљују спонзорима, као и свима који су подржали организацију конференције. Њихова искрена подршка и посвећеност циљевима скупа били су од пресудног значаја за његову реализацију.

Верујемо да у времену које долази можемо заједничким напорима да пружимо значајан допринос даљим истраживањима у овој научној области од изузетног значаја.

Садржај

Радови по позиву

Излагања

Информатичка обрада јужнословенских језика

Дигитални корпуси јужнословенских језика

Ресурси за обраду јужнословенских језика

Технологије обраде јужнословенских језика

Граматика и лексика јужнословенских језика у контексту обраде природних језика

Вештачка интелигенција, језички модели и обрада јужнословенских језика

Дигитална хуманистика

ЈуДиг радионице

Радови по позиву

Агата Савари

Професор рачунарства, Универзитет Париз-Сакле
LISN (Интердисциплинарна лабораторија дигиталних наука)
Универзитетски институт за технологију Орсеја
https://perso.lisn.upsaclay.fr/savary/

Агата Савари је професор рачунарства на Универзитету Париз-Сакле у Француској. Магистрирала је на Универзитету у Варшави у Пољској, докторирала на Универзитету Марн-ла-Вале у Француској и стекла хабилитицију на Универзитету у Туру у Француској. Вишејезичном обрадом природних језика се бави три деценије. Њени домени интересовања обухватају обраду природног језика (НЛП), универзалистичко моделирање идиоматичности у језику које се може применити у обради природних језика, као и изградња језичких ресурса и алата за идентификацију вишечланих израза, препознавање именованих ентитета и разрешавање кореференци.

Агата Савари председава акцијом CA21167 COST UniDive (Универзалност, разноликост и идиосинкразија у језичким технологијама, 2022-2026), која окупља преко 300 чланова из 37 земаља.

Такође је председавала акцијом IC1207 COST PARSEME (Парсирање и вишечлани изрази, 2013-2017), била је изабрана представница секције Multiword Expressions у оквиру SIGLEX-у, ко-уредник серије књига „Фразеологија и вишечлани изрази" у издањуLanguage Science Press-а и координатор Дагстухл семинара "Универзалије лингвистичке идиосинкразије у вишејезичкој рачунарској лингвистици". Објавила је 20 рецензираних радова у часописима, 40 радова у зборницима са конференција, 20 радова у зборницима са радионица и 10 поглавља у књигама.

Аутоматска идентификација вишечланих израза – недавно постигнути напредак и перспективе

Изрази са више речи, као што су на енглеском "all of a sudden", "hot dog", или "hot dog", представљају комбинације речи које показују идиосинкратично понашање на лексичком, морфолошком, синтаксичком, семантичком, прагматичком или статистичком нивоу. Њихова семантичка некомпозитност је њихова најистакнутија карактеристика која може представљати озбиљан изазов у семантички оријентисаним задацима обраде природних језика.

Један од начина хватања у коштац са овим изазовом је да се вишечлани изрази идентификују у текућем тексту пре него што се на њега примени жељена обрада.

У идентификацију вишечланих израза уложено је много напора, а посебно у оквиру заједничког задатка аутоматске идентификације глаголских вишечланих израза који се решавао под окриљем COST акције PARSEME. Сумираћу главне налазе ових заједничких задатака и истаћи посебно оне особине вишечланих израза које њихову идентификацију чине изазовном, чак и када се користе методе дубоког учења.

Такође ћу се дотакнути најновијих изазова и могућности обраде вишечланих израза код примене на генеричке задатке обраде природних језика, као што су неуронско машинско превођење или интерпретација неуронских модела.

Светла Коева

Професор при Бугарској академији наука
Директор Института за бугарски језик
Председавајући Катедре за рачунарску лингвистику
http://dcl.bas.bg/news/svetla-koeva/

Др Светла Коева је професор рачунарске лингвистике и шеф одељења за рачунарску лингвистику при Институту за бугарски језик Бугарске академије наука. Њена истраживачка интересовања су у области рачунарске лингвистике и формалног описа језика: морфологије и синтаксе, лексичко-семантичких мрежа и онтологија. Она је водећи истраживач за развој различитих језичких ресурса за бугарски језик као што су: бугарски вордент, бугарски национални корпус, ланац за обраду текста итд.

Светла Коева је објавила 5 књига и преко 200 истраживачких публикација. Водила је бројне истраживачке пројекте међу којима су тренутно у току: Процена читалачке писмености и разумевања ученика нижих разреда у Бугарској и Италији и Обогаћивање семантичке мреже ворднет концептуалним оквирима. Успешно завршени пројекти у последње две године су: Европска лексикографска инфраструктура, вишејезични мултимедијални корпус (MIC 21), Онтологија стативних ситуација у језичким моделима, Прирећени вишејезични ресурси за CEF.AT. Светла Коева од 2013. води пројекат Писана реч остаје. Пишите исправно! чији је циљ унапређење проучавања бугарског језика. Такође је шеф Националног центра компетенција у Европској језичкој мрежи.

Била је директорка Института за бугарски језик од 2012. до 2021. године, а од 2021. је председница Истраживачког савета Института за бугарски језик. Светла Коева је и главни уредник Годишњих радова Института за бугарски језик као и недељног издања Писана реч остаје. Пишите исправно!. Светла Коева је награђена са пет награда Националног фонда за науку при Министарству просвете и науке и Бугарске академије наука.

(Рачунарско-)лингвистичко истраживање засновано на језичким моделима: између добити и ризика

Област рачунарске лингвистике је претрпела велике концептуалне промене, крећући се од симболичких техника ка машинском учењу и дубоком учењу. Велики језички модели ((LLMs) сада замењују многе традиционалне технологије обраде природних језика у различитим областима примене као што су одговарање на питања, сажимање текста, поједностављење текста и препознавање именованих ентитета. Штавише, показује се да су језички модели способни за анализу података у оквиру датог теоријског оквира, за истраживање семантике оквира и лингвистичке студије засноване на корпусима аутоматским означавањем текстова специфичним лингвистичким информацијама.

Циљ овог предавања је разматрање изазова које коришћење великих језичких модела у (рачунарским) лингвистичким истраживањима поставља. Предавање ће се ослањати на постојеће и текуће студије. Истражићемо предности, изазове, ограничења и потенцијалне ризике које употреба великих језичких модела у истраживању носи, са посебним нагласком на јужнословенске језике у поређењу са енглеским. Циљ нам је да кроз ову анализу пружимо увид у примену великих језичких модела у рачунарској лингвистици и идентификујемо правце даљих истраживања и побољшања.

Руслан Митков

Професор рачунарства и комуникација, Факултет рачунарства и комуникација, Универзитета Ланкастер
Професор рачунарске лингвистике и језичког инжењерства
Директор Erasmus Mundus програма за технологије превођења и интерпретирања (EM TTI)
Извршни уредник часописа 'Natural Language Processing' (Cambridge University Press)
Програмски председник серије конференција RANLP
https://wp.lancs.ac.uk/mitkov

Проф. др Руслан Митков је професор рачунарства и комуникација на Универзитету Ланкастер, једном од 10 најбољих универзитета у Великој Британији. Пре него што се придружио Универзитету Ланкастер, проф. Митков је радио на Универзитету у Вулверхемптону где је створио и водио међународно признату истраживачку групу из области рачунарске лингвистике, а такође је био и директор Истраживачког института за обраду информација и језика, као и директор Лабораторије за одговорну дигиталну хуманистику. Проф. Митков је такође истакнути професор на Универзитету Аликанте, Шпанија.

Професор Митков ради у области обраде природних језика (ОПЈ), рачунарске лингвистике, корпусне лингвистике, машинског превођења, технологије превођења и сличних области од раних 1980-их. Док је професор Митков најпознатији по свом значајном доприносу у области разрешавања анафора, аутоматског генерисања тестова са вишеструким избором и нове генерације система заснованих на преводилачким меморијама, његова високо цитирана библиографија (више од 320 публикација укључујући 15 књига, 35 чланака у часописима и 35 поглавља у књигама) покрива, између осталог, и теме као што су рачунарска фразеологија, машинско превођење, обрада природног језика за сметње у комуникацији, аутоматска сумаризација, рачунарски подржана обрада језика, анотирање корпуса, двојезична екстракција термина, аутоматска идентификација когната и "лажних пријатеља", проучавање преводних универзалија засновано на корпусима методама обраде природних језика и поједностављење текста. Његова недавна истраживања укључују употребу дубоког учења, великих језичких модела и вештачке интелигенције у обради природних језика, технологији превођења, лингвистици и истраживању језика уопште. Проф. Митков није познат само по својим оригиналним истраживачким резултатима са високим научним утицајем, већ је познат и по својој визионарским и иновативним примењеним истраживањима која настоје да унапреде ефикасност рада различитих професија (наставника, преводилаца и интерпретатора) и да побољшају квалитет живота (особа са посебним потребама).

Проф. др Руслан Митков је професор рачунарства и комуникација на Универзитету Ланкастер, једном од 10 најбољих универзитета у Великој Британији. Пре него што се придружио Универзитету Ланкастер, проф. Митков је радио на Универзитету у Вулверхемптону где је створио и водио међународно признату истраживачку групу из области рачунарске лингвистике, а такође је био и директор Истраживачког института за обраду информација и језика, као и директор Лабораторије за одговорну дигиталну хуманистику. Проф. Митков је такође истакнути професор на Универзитету Аликанте, Шпанија.

Професор Митков ради у области обраде природних језика (ОПЈ), рачунарске лингвистике, корпусне лингвистике, машинског превођења, технологије превођења и сличних области од раних 1980-их. Док је професор Митков најпознатији по свом значајном доприносу у области разрешавања анафора, аутоматског генерисања тестова са вишеструким избором и нове генерације система заснованих на преводилачким меморијама, његова високо цитирана библиографија (више од 320 публикација укључујући 15 књига, 35 чланака у часописима и 35 поглавља у књигама) покрива, између осталог, и теме као што су рачунарска фразеологија, машинско превођење, обрада природног језика за сметње у комуникацији, аутоматска сумаризација, рачунарски подржана обрада језика, анотирање корпуса, двојезична екстракција термина, аутоматска идентификација когната и "лажних пријатеља", проучавање преводних универзалија засновано на корпусима методама обраде природних језика и поједностављење текста. Његова недавна истраживања укључују употребу дубоког учења, великих језичких модела и вештачке интелигенције у обради природних језика, технологији превођења, лингвистици и истраживању језика уопште. Проф. Митков није познат само по својим оригиналним истраживачким резултатима са високим научним утицајем, већ је познат и по својој визионарским и иновативним примењеним истраживањима која настоје да унапреде ефикасност рада различитих професија (наставника, преводилаца и интерпретатора) и да побољшају квалитет живота (особа са посебним потребама).

Професор Митков је аутор монографије Разрешавање анафора (Longman) и једини уредник Оксфордског приручника за рачунарску лингвистику (Oxford University Press) који је проглашен за најуспешнији Оксфордски приручник и чије је друго, значајно ревидирано издање објављено јуна 2022. Текући престижни пројекти укључују позицију извршног уредника часописа Natural Language Processing (раније Journal of Natural Language Engineering) који издаје Cambridge University Press, главног уредника серије књига Обрада природног језика издавача John Benjamins и уредника консултанта за публикације Охфорд Университу Пресс из области рачунарске лингвистике.

Проф. Митков је био позвани предавач на више од 240 међународних конференција (28 уводних предавања само у 2024. години), а председавао је или и даље председава на више од 70 конференција из области обраде природног језика, технологије превођења и примењене лингвистике. Уредио је више од 15 томова које су објавили Спрингер (Springer) и Џон Бенџаминс (John Benjamins).

Митков је осмислио први и једини Еразмус Мундус мастер програма из технологија за превођење и интерпретацију, чији је тренутно директор. То је иновативан и инспиративан програм чији је фокус подједнако на истраживању и на бизнису; водеће светске компаније које се баве језичким и преводилачким технологијама учествују као придружени партнери.

Проф. Митков је био спољшни испитивач бројних докторских дисертација и наставних планова и програма у Великој Британији и иностранству, укључујући мастер програме који се односе на обраду природног језика, рачунарску лингвистику, дигиталне хуманистичке науке, превођење и технологију превођења.

Руслан Митков је магистрирао на Хумболт универзитету у Берлину, докторирао на Техничком универзитету у Дрездену и радио је као професор истраживач на Институту за математику Бугарске академије наука у Софији. Био је стипендиста Фондације Алекандер вон Хумболдт, из Немачке, стипендиста програма Марија Кири, уважени гостујући професор на Универзитету Франш-Конте у Безансону, у Француској и уважени гостујући истраживач на Универзитету у Малаги, у Шпанији.

Руслан Митков је потпредседник AsLing, међународног удружења за промоцију језичких технологија. У септембру 2022. године реномирани Национални одбор медицинских испитивача из САД-а уручио је проф. Миткову сертификат о истакнутој сарадњи која је трајно утицала на стратешко планирање и доношење одлука ове америчке организације и на њихово коришћење техника обраде природног језика за процењивање у последњих 17 година. Као признање за изузетна професионална и истраживачка достигнућа, проф. Митков је три пута одликован звањем почасног доктора.

Еволуција обраде природног језика: од правила преко неуронских мрежа до генеративне АИ. Шта доноси будућност?

Обрада природног језика (ОПЈ) пролази кроз динамичне и неслућене промене. Одувек смо знали да обрада природних језика није магична технологија јер су њени резултати нису били ни близу 100% тачни, област технологије језика и превођења се мења. Методе дубоког учења, а затим и велики језички модели, муњевито су освојили свет. Ово забавно предавање које се лако прати настојаће да осветли будућност обраде природних језика у ери вештачке интелигенције (ВИ).

Предавање ће скицирати историју обраде природних језика и машинског превођења и дати преглед најновијих достигнућа заснованих на дубоком учењу и великим језичким моделима (LLM). Затим ће бити дат критички осврт на примену великих језичких модела у обради природног језика и машинском превођењу, кроз приказ недавних истраживања предавача у којима упоређује велике језичке моделе, дубоко учење и приступе засноване на правилима за одабране задатке обраде природног језика и конкретне примене.

Приказане студије случаја ће послужити као основа за наставак дискусије о будућности обраде природног језика. Предавач ће нагласити да није видовит, већ да ће на основу свог искуства у овој области покушати да предвиди вероватну будућност вештачке интелигенције поређујући је са људском интелигенцијом а користећи језик као опитни полигон.

Цветана Крстев

Професор информатике на Филолошком факултету Универзитета у Београду (у пензији)
Председник Друштва за језичке ресурсе и технологије (JePTex)
https://poincare.matf.bg.ac.rs/~cvetana/

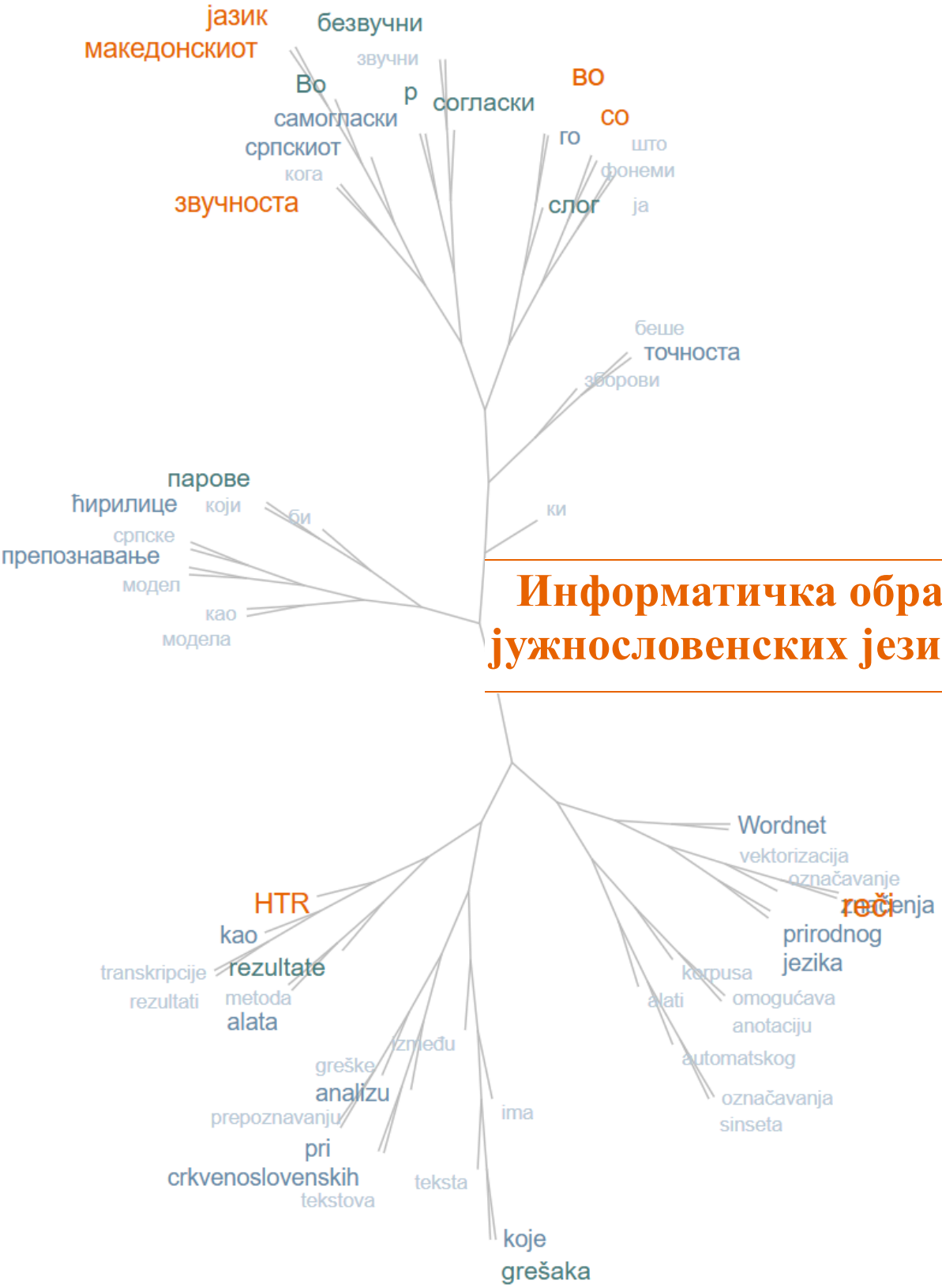
Проф. др. Цветана Крстев је професор информатике на Катедри за библиотекарство и информатику Филолошког факултета Универзитета у Београду, у пензији. Научна област њених истраживања су језичке технологије. Објавила је једну књигу и више од 200 научних радова, већином везаних за обраду природног језика, конкретно за развој језичких ресурса и њихову примену. Развила је Српски морфолошки електронски речник и један је од кључних сарадника у развоју српског вордента, једнојезичних корпуса српског језика и паралелних двојезичних и вишејезичких корпуса, потом система за препознавање именованих ентитета у тексту на српском језику, као и система за аутоматско кориговање текста, за различите трансформације текста, интеракцију корпуса и лексикона, као и многих других језичких ресурса и алата.

Учествовала је у неколико међународних и домаћих пројеката везаних за развој језичких ресурса и технологија. Председник је и један од оснивача Друштва за језичке ресурсе и технологије (JePTex). Такође је шеф Националног центра компетенција у Европској језичкој мрежи. Била је члан управног одбора и активни учесник акција IC1207 COST PARSEME (Парсирање и вишечлани изрази, 2013-2017), D-Reading (Удаљено читање за историју европске књижевности, 2017-2021) а тренутно води активности као члан управног одбора акције CA21167 COST UniDive (Универзалност, разноликост и идиосинкразија у језичким технологијама, 2022-2026).

Однос „нас“ и „других“ у корпусу SrpELTeC добијен старим добрим методама

SrpELTeC корпус је део велике вишејезичне збирке романа писаних у периоду 1840-1920. Садржи 120 романа написаних оригинално на српском језику. Због својих прецизних анотација већ је коришћен за истраживања у различитим доменима: лингвистичким, филолошким и културолошким. У овом истраживању покушаћемо да утврдимо како се однос „нас“ (Срби по нацији или држављанству) и „други“ одражава у романима писаним у другој половини 19. и почетком 20. века. Користићемо анотације које су унете ручно, полуаутоматски или аутоматски, при чему се последње две ослањају на свеобухватне лексичке ресурсе и алате засноване на њима. Покушаћемо да покажемо да добре старе методе, као што су читање и бројање, још увек могу дати неке занимљиве резултате.

Излагања



Информатичка обрада јужнословенских језика

Saša Petalinkar
Univerzitet u Beogradu, Jerteh
E-mail: sasa5linkar@gmail.com

Automatsko označavanje sinseta za Srpski Wordnet: razvoj, predobrada i procena izvodljivosti

Ovo istraživanje predstavlja razvoj anotatora za Srpski Wordnet, osmišljenog da dodeli odgovarajuće sinsete svakoj reči u datom tekstu. Osnovni okvir je zasnovan na spaCy pipeline-u poboljšanom BERT Jerteh 355 modelom, koji omogućava različite zadatke obrade prirodnog jezika, uključujući lematizaciju. Prilagođeni sloj je integrisan radi omogućavanja označavanja sinseta.

Zbog složenosti lematizacije n-grama, ovo istraživanje se fokusira na označavanje unigrama sa njihovim odgovarajućim sinsetima. Glavni izazov koji se razmatra je razlučivanje značenja reči za reči sa više značenja. Ova studija ima za cilj da proceni izvodljivost automatskog rešenja korišćenjem ugrađivanja i vektorske sličnosti između reči i opisa sinseta, ili neophodnost ručne anotacije za izgradnju korpusa za dalju obuku.

Korist od Wordnet anotatora je u tome što omogućava tačno i brzo označavanje značenja reči, što je od suštinskog značaja za obradu prirodnog jezika i brojne aplikacije u oblasti veštačke inteligencije. Wordnet pruža strukturiranu bazu podataka koja povezuje reči sa njihovim značenjima i odnosima, čime se unapređuje razumevanje konteksta i semantike.

Kao predobrada, vrši se vektorizacija svih definicija iz Srpskog Wordneta. U odgovarajućem sloju se uzima lema dobijena u prethodnim slojevima spaCy pipeline-a i pretragom se nalaze svi sinseti koji sadrže tu reč. Ukoliko postoji jedan ili nijedan sinset, on se automatski dodaje; ako ih ima više, vrši se vektorsko poređenje između vektora reči i već vektorizovanih definicija. U istraživanju se otkrivaju koji tipovi vektorizacije su mogući i kakve rezultate daju.

Preliminarni rezultati su pokazali da vektorizacija korišćenjem SrpCNNER-a daje sledeće rezultate kada je primenjena na dvadeset rečenica paralelnog englesko-srpskog korpusa ručno anotiranog Wordnetom: 58 tačnih i 87 netačnih.

Osim automatskog označavanja, razvijaju se alati koji pomažu anotatorima predlažući moguće sinsete, što omogućava ručnu anotaciju i izgradnju korpusa za dalju obuku. Ovi alati omogućavaju procenu koliko vektorsko poređenje može postići dobre rezultate, ili da li je neophodno dalje poboljšanje kroz ručnu anotaciju.

Ovi nalazi će doprineti daljem razvoju alata za srpski jezik, pružajući uvide u potencijale i ograničenja automatskog označavanja sinseta i razlučivanja značenja reči. Očekuju se značajne implikacije za unapređenje razumevanja i obrade prirodnog jezika na srpskom jeziku.

Ključne reči: *automatsko označavanje sinsetova, Srpski Wordnet, vektorizacija reči, obrada prirodnog jezika, disambiguacija sinstetova, BERT model*

Владимир Поломац, Тамара Лутовац Казновац, Марко Милошевић, Ана Марија Павловић
Univerzitet u Kragujevcu, FILUM, Katedra za srpski jezik
E-mail: {v.polomac|tamara.kaznovac| marko.milosevic| ana.marija.pavlovic |@filum.kg.ac.rs

Модели за аутоматско препознавање старе српске штампане и рукописне ћирилице: тренутно стање и будући задаци

У раду се даје преглед модела за аутоматско препознавање старе српске ћирилице који су развијени у софтверској платформи Транскрибус у оквиру међународног билатералног пројекта *Креирање АИ модела за аутоматску обраду српских средњовековних рукописа* Катедре за српски језик Универзитета у Крагујевцу и Института за славистику Универзитета у Фрајбургу. Прегледом су обухваћени 1) велики енерички модел за препознавање српскословенске штампане и рукописне (уставне, полууставне и брзописне) ћирилице XII–XVIII века, 2) генерички модел за препознавање српскословенске и српске полууставне и брзописне ћирилице у рукописима Гаврила Стефановића Венцловића (XVIII век), и 3) генерички модел за препознавање српске дипломатичке ћирилице XIII–XVI века. У раду су представљене квантитативне и квалитативне перформансе наведених модела, као и могући правци њиховог даљег развоја.

Кључне речи: *аутоматско препознавање текста, Транскрибус, вештачка интелигенција, машинско учење, стара српска ћирилица*

Dr. Anna Jouravel, Martin Meindl, Prof. Dr. Achim Rabus, Elena Renje
University of Freiburg
E-mail: {anna.jouravel, martin.meindl, achim.rabus, elena.renje}@slavistik.uni-freiburg.de

Kreiranje multistratalnog digitalnog okruženja za analizu crkvenoslovenskih rukopisa

U ovom izlaganju govorićemo o različitim digitalnim metodama za analizu staroslovenskih i crkvenoslovenskih tekstova istočnoslovenske i južnoslovenske redakcije. Za analizu koristimo multistratalni, a ne monostratalni pristup, uz primenu različitih alata i metoda. Verujemo da ovakva strategija daje validnije i robusnije rezultate, jer je manje verovatno da će skup kombinovanih metoda sadržati one greške koje prate primenu samo jednog alata.

Ove metode deo su kvantitativnog lanca koji počinje automatskom transkripcijom (staro)crkvenoslovenskih rukopisa, a nastavlja se analizom ovih transkripata, bez potrebe za manuelnim ispravljanjem grešaka koje nastaju primenom programa za prepoznavanje teksta napisanog rukom (HTR). Kvantitativni lanac obuhvata i nekoliko koraka koji prethode obradi ili joj slede, uključujući automatsku transkripciju, tokenizaciju, raščlanjivanje, anotaciju (tagiranje) i statističku analizu. Zasniva se na upotrebi HTR programa kao što je platforma Transkribus, zatim na softveru zasnovanom na pretraživačima i aplikacijama kao što su

UDPipe, Stanza, Voyant Tools i AntConc, kao i na statističkim metodama implementiranim u programskim jezicima poput R-a i Python-a.

Rabus (2019) je pokazao da HTR-alati mogu automatski da transkribuju velike količine crkvenoslovenskih tekstova sa procentom greške pri prepoznavanju karaktera (CER) od 4% ili manje, što je rezultat koji je u skladu sa performansama modela koji smo koristili u našim eksperimentima. Međutim, kako se pri prepoznavanju leksičkih jedinica najveće frekvencije javlja mnogo manji broj grešaka, predlažemo da rezultati transkripcije služe kao stabilna osnova za pojedine vrste kvantitativnih analiza, s tim što smatramo da velika količina transkribovanog teksta ima veću težinu od relativno malog broja uočenih grešaka. Koristeći mali fragment teksta izvršili smo paralelnu kvantitativnu analizu najčešćih grešaka koje nastaju prilikom primene HTR-alata. U skladu sa prethodno dobijenim rezultatima (Rabus 2019, Burlacu & Rabus 2021), utvrdili smo da su greške pri prepoznavanju eksponentnih slova i razmaka između reči dva najčešća uzroka grešaka koje se javljaju prilikom transkripcije. Ova analiza nam je omogućila da bolje procenimo efekat pogrešne transkripcije i na naše rezultate.

Pristup koji se zasniva na upotrebi multistratalnih, mešovitih metoda ne samo da obezbeđuje da rezultati dobijeni iz HTR- podataka budu logični i smisleni, već se pokazao uspešnim i pri analizi crkvenoslovenskih tekstova (Rabus & Petrov 2023).

Ključne reči: *mešovite metode analize, crkvenoslovenski jezik, statistika, HTR*

Literatura

[1] Burlacu, C. & Rabus, A. (2021): Digitising (Romanian) Cyrillic using Transkribus: new perspectives. Diacronica, 14, art. A196. P. 1-9.
[2] Rabus, A. (2019): Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. Scripta & e-Scripta, 19. P. 9-32.
[3] Rabus, A. & Petrov, I. N. (2023): Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach. Scando-Slavica, 69.1. P. 25-38.

Катерина Здравкова, Јана Кузманова
Факултет за информатички науки и компјутерско инженерство, Скопје
E-mail: {katerina.zdravkova; jana.kuzmanova}@finki.ukim.mk

Поделба на македонските и српските зборови на слогови по звучност

Фонетски, слоговите претставуваат единици на говорните звуци, додека фонолошки, тие се единици за поставување на нагласокот. Според „Принципот за секвенционирање на звукот“, звучноста во еден слог го достигнува својот максимум во нуклеусот на слогот, а потоа паѓа кон границите. Направени се неколку обиди за поделба на македонските и српските зборови на слогови. Точноста на македонскиот експеримент не беше оценета на конкретен корпус, додека српската поделба надминува 98%. Пристапот заснован на правила е прилично сложен во споредба со поделбата заснована на звучност што ја предложивме за македонскиот, а сега ја прошируваме и за српскиот јазик.

Звучноста на македонските фонеми зависи од нивната основна класификација: самогласки (тежина 12), сонанти (4), звучни (2) и безвучни согласки (1). Кога звучникот р (латинска транскрипција: r) е меѓу две согласки, тој станува носител на слог и затоа

неговата звучност е поголема, првично 6. Две едноподруги самогласки се одделуваат со фиктивна согласка FC чија тежина е 0.

Звучноста на српските фонеми ги опфаќа следниве класи: самогласки (12), слоготворно р (8), експлозивни звучни согласки (4), експлозивни безвучни и безвучни фрикати (3), безвучни фрикати (2), звучни африкати (2) и безвучни африкати (1).

Нуклеусите на слогот во обата јазика се петте самогласки. Во македонскиот јазик, слоготворно е р кога се појавува во средината на консонантска група (крст, вр-ста, пр-вен-ство) или на крајот од зборот (ма-са-кр). Во српскиот јазик, освен сонантот р (тврд, цр-ве-но, тр-ка) и сонантите л и н можат да станат слоготворни (на пример, би-ци-кл, Вл-та-ва, Њу-тн). Нуклеусите за секој слог се одредуваат со пресметување на разликата на звучноста на тројката фонеми: тековната и нејзиниот лев и десен сосед.

Одредувањето на границите на слоговите зависи од монотоното неопаѓање или опаѓање на звучноста. Во македонскиот јазик, секогаш кога звучноста на две согласки е неопаѓачка, тие припаѓаат на два соседни слога. Во српскиот, кога звучноста е неопаѓачка, тогаш и двете согласки се дел од вториот слог.

Во македонскиот јазик, точноста на основниот алгоритам беше прилично мала, првенствено заради наставките ски, ство и ствен и нивните флексии, кои треба да останат во еден слог. Вклучувајќи го ова правило, постигнавме точност од 95.60%, оценета на корпус од повеќе од 1000 зборови. Овој исклучок влијаеше неповолно врз поделбата на именките: гус-ки, мас-ки, прас-ки, во кои ски не е морфема. Врз основа на примерок од повеќе од 3000 српски зборови, точноста на основниот алгоритам беше 97.59%. Со промена на звучноста на неслоготворното р на 6, точноста достигна 98.54%, надминувајќи ја точноста на пристапот што се темели врз правила.

Пристапот што го предложивме е исклучително едноставен и во исто време, многу ефикасен. Имаме намера дополнително да го подобриме со земање предвид на зборовната група за македонскиот јазик и со вклучување на исклучоците за српскиот, со надеж дека ќе достигнеме точност над 99%.

Кључне речи: *македонски, српски, звучност на фонемите, поделба на слогови*

Литература

[1] G. Clements: The sonority cycle and syllable organization, Phonologica,. 63-76, 1988.
[2] M. Mitreska, K. Zdravkova: Syllable and Morpheme Segmentation of Macedonian Language. Proceeding of 46th MIPRO, 1113-1118. IEEE, 2023.
[3] A. Kovač, M. Marković M: A Mixed-principle Rule-based Approach to the Automatic Syllabification of Serbian. Contributions to Contemporary History / Prispevki za Novejšo Zgodovino. 2019.

Данило Алексић
Катедра за српски језик са јужнословенским језицима
Филолошки факултет, Универзитет у Београду
E-mail: danilo.aleksic@fil.bg.ac.rs

Нове унапређене верзије програма „Ка минималним паровима”

Кратки програм на Python-у Ка минималним паровима (КаМП), први пут представљен у Алексић, Шандрих 2021, замишљен је првенствено као алат помоћу којег наставници српског као страног језика могу прикупити парове речи за вежбе изговора. КаМП из латиничког српског улазног корпуса у Unicode-у аутоматски ексцерпира сегменталне минималне парове, нпр. lak ~ luk (лак ~ лук), и парове сродне сегменталним минималним паровима (тј. парове који би били сегментални минимални парови када би се игнорисала прозодија), нпр. gel ~ gen (гел ~ ген). У Алексић, Мркела 2022 представљене су брже варијанте КаМП-а са побољшаним сортирањем и са суплементарним модом, КаМП 2 и КаМП 2.1. Овом приликом представљају се још брже варијанте КаМП-а, КаМП 2.2 и КаМП 2.3. Сегменталне минималне и сродне парове који се разликују по поднискама "dž" и "đ" из корпуса POL.xml (без разликовања великих и малих слова) у Python-у 3.12.4 КаМП ексцерпира за 122,66, КаМП 2 за 74, КаМП 2.1 за 63,84, КаМП 2.2 за 56,51, а КаМП 2.3 за 52,74 секунде (просеци 20 узастопних мерења по верзији). Убрзање је углавном постигнуто коришћењем класа уместо речника на одређеним местима у коду. Како се види, КаМП 2.2 и КаМП 2.3 покренути на великим корпусима (POL.xml броји око 117.900.900 речи /Алексић, Шандрих 2021: 575/) крајњем кориснику омогућују приметну уштеду у времену у односу на КаМП 2.1.

Један податак из Python 2024 („Attribute lookup speed can be significantly improved . . .”) водио је ка очекивању да ће верзија КаМП-а у којој се користе класе са класовном променљивом __slots__ (КаМП 2.2) бити бржа од верзије КаМП-а у којој се користе класе без те класовне променљиве (КаМП 2.3), али су малопре наведени резултати мерења показали супротно. Будући да је ова релативна спорост „слотованих” класа донекле сумњива, потребно је приказати дефиниције са класовном променљивом __slots__ коришћене у КаМП-у 2.2:

```
class Dictionary:
    tokens_ = tokenization()
    __slots__ = tokens_

    def __init__(self):
        for word in self.tokens_:
            setattr(self, word, word.casefold())

class Dictionary2:
    __slots__ = tuple_2[1]

    def __init__(self):
        for tuple_ in tuple_2[0]:
            if not hasattr(self, tuple_[0]):
                setattr(self, tuple_[0], {})
            getattr(self, tuple_[0])[
                tuple_[1]] = tuple_[2]
```

```
class Dictionary3:
   __slots__ = tuple_2[1]

   def __init__(self):
      for tuple_ in tuple_2[0]:
         if not hasattr(self, tuple_[0]):
            setattr(self, tuple_[0], {})
         getattr(self, tuple_[0])[
            tuple_[1]] = tuple_[2]

def excerpt(letter):
   return (word
         for word in dictionary_from_corpus.__slots__
         if letter in getattr(dictionary_from_corpus, word))
```

Наведени резултати мерења могу бити од користи програмерима, јер показују да у Python-у 3.12.4 (1) класе могу бити брже од речника и да (2) класе без слотова могу бити брже од класа са слотовима (ако су у КаМП-у 2.2 слотови употребљени правилно, тј. са највећом могућом ефикасношћу).

Кључне речи: *минимални парови, фонетика, фонологија, обрада природног језика, корпусна лингвистика, Python*

Литература

[1] Aleksić, Danilo, and Lazar Mrkela. 2022. “The Enhanced Versions of the Program “Ka Minimalnim Parovima” (Towards Minimal Pairs).” *Infotheca* 22 (1): 7–31. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2022.22.1.1_en.

[2] Python 2024: Python 3.12.5 documentation. Accessed August 22, 2024. https://docs.python.org/3/.

[3] Алексић, Данило, and Бранислава Шандрих. 2021. “Аутоматска ексцерпција парова речи за учење изговора у настави српског као страног језика.” *Српски језик* 26 (1): 567–584. https://doi.org/10.18485/sj.2021.26.1.3.

Дигитални корпуси јужнословенских језика

Jelena Redli
Univerzitet u Novom Sadu – Filozofski fakultet
E-mail: redli@ff.uns.ac.rs

Značaj digitalnog korpusa i jezičkih alata u jezičkoj analizi forenzičkih tekstova na srpskom jeziku

Forenzička lingvistika (FL) predstavlja interdisciplinarnu oblast koja objedinjuje jezičke, pravne i digitalne tehnologije kako bi unapredila razumevanje jezičkih fenomena u pravnom i kriminalističkom kontekstu. Ona već skoro jedan vek pronalazi primenu u raznim oblastima poput identifikacije autora, dokazivanja povrede autorskih prava, kao i otkrivanja mogućih kriminalnih radnji. Za svaku FL aktivnost, od suštinskog je značaja postojanje odgovarajućih korpusa, koji će se koristiti za obuku sistema tokom njihove izrade. Njihova upotreba u forenzičkim istraživanjima postaje neizostavan alat u globalnom okruženju, ali srpski jezik ostaje značajno neistražen u ovom polju zbog nedostatka specijalizovanog korpusa forenzičkih tekstova.

Ovaj rad istražuje potencijal kreiranja i primene korpusa forenzičkih tekstova na srpskom jeziku, uz naglasak na tome kako bi takav korpus mogao revolucionarizovati analizu autentičnosti teksta, identifikaciju autora i razrešavanje pravnih sporova putem jezičkih dokaza. Rad predlaže metodologiju za formiranje korpusa koji bi obuhvatio različite tipove forenzičkih tekstova, uključujući policijske izveštaje, pravne dokumente, preteće poruke, oproštajna pisma i druge dokumente relevantne za pravosudni kontekst. Biće predstavljene i strategije za primenu naprednih jezičkih alata za efikasno izvođenje jezičkih analiza, uključujući softverske alate za automatsku analizu teksta, kao i mogućnosti za dubinsku analizu leksičke frekvencije, sintakse, stilskih markera, jezičkih ograda, rodnolekata i sl.

Osim teorijskog doprinosa, rad ima za cilj da praktično demonstrira kako bi konkretna implementacija ovakvog korpusa unapredila preciznost forenzičkih analiza u srpskom kriminalističko-pravnom kontekstu. U radu se razmatraju i potencijalni izazovi u razvoju korpusa, uključujući tehničke, pravne i zakonske aspekte, kao i moguće prepreke u prikupljanju i obradi osetljivih podataka. Poseban akcenat stavlja se na neophodnost pridržavanja etičkih i pravnih smernica tokom razvoja ovakvih resursa.

Na kraju, rad daje praktične preporuke za buduća istraživanja i implementaciju u praksi, i nudi osnovu za dalji razvoj forenzičke lingvistike u Srbiji. Takođe, ističe značaj interdisciplinarne saradnje između lingvista, pravnika, policije i IT stručnjaka u cilju razvijanja velikih jezičkih resursa koji mogu služiti kao ključni alati u forenzičkim istraživanjima.

Ključne reči: *forenzička lingvistika, srpski jezik, pravni jezik, digitalni korpus, jezički alati, forenzička analiza teksta, identifikacija autora*

Literatura

[1] Blackwell, S. (2009). Why forensic linguistics needs corpus linguistics. *Comparative Legilinguistics*, *1*, 5–19.
[2] Chaski, C. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. International Journal of Digital Evidence 4(1), 1–14.
[3] Cotterill, J. (2010). How to use corpus linguistics in forensic linguistics. U A. O'Keefe i M. McCarthy (ur.), The Routledge Handbook of Forensic Linguistics, 578–590.
[4] Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech, Language and the Law*, *1*(1), 27–43.

[5] Goźdź-Roszkowski, S. (2021). Corpus linguistics in legal discourse. *International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique*, *34*(5), 1515–1540.
[6] Hercigonja-Szekeres, M., Sikirica, N., & Popović, I. (2012). Statistička analiza tekstnih podataka. *In medias res: časopis filozofije medija*, *1*(1), 79–91.
[7] Lalić, A. (2024). Elektronski korpus sms poruka na bosanskom jeziku (Halid Bulić, Elma Durmišević, Azra Hodžić-Čavkić, Enisa Bajraktarević, Azra Ahmetspahić-Peljto, Belmin Šabić, Sarajevski korpus SMS poruka na bosanskom jeziku, Univerzitet u Sarajevu–Filozofski fakultet, Sarajevo, 2023). *Društvene i humanističke studije*, *9*(1 (25)), 1187–1190.
[8] Longhi, J. (2021). Using digital humanities and linguistics to help with terrorism investigations. *Forensic Science International*, *318*, 110564.
[9] Vitas, D., Krstev, C., Obradovic, I., Popovic, L., & Pavlovic-Lazetic, G. (2003, November). An overview of resources and basic tools for processing of Serbian written texts. In *Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*.
[10] Vitas, D., Krstev, C., Obradović, I., Popović, L., & Pavlović-Lažetić, G. (2003, November). Processing Serbian written texts: An overview of resources and basic tools. In *Workshop on Balkan Language Resources and Tools* (Vol. 21, pp. 97–104).
[11] Wright, D. (2017). Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International journal of corpus linguistics*, *22*(2), 212–241.
[12] Wright, D. (2020). Corpus approaches to forensic linguistics: Applying corpus data and techniques in forensic contexts. In *The Routledge handbook of forensic linguistics*, 611–627. Routledge.

Milica Dinić Marinković
Univerzitet u Beogradu – Filološki fakultet, Katedra za opštu lingvistiku
E-mail: milica.dinic.marinkovic@fil.bg.ac.rs

Milena Oparnica
Istraživačko-razvojni institut za veštačku inteligenciju Srbije
E-mail: milena.oparnica@ivi.ac.rs

DeciKo – korpus knjiga za decu ranog i predškolskog uzrasta. Trenutno stanje i izazovi

Uticaj (ne)upražnjavanja posrednog čitanja (od najranije dobi) na jezički i kognitivni razvoj dece neupitno je potvrđen u mnogobrojnim empirijskim istraživanjima (za pregled v. Nation i sar., 2022). Uprkos tome, uzročna veza između izlaganja dece jeziku knjige i pozitivnog merljivog ishoda na nivou komunikativne kompetence nije u potpunosti sagledana niti adekvatno objašnjena. Razlog za ovaj nedostatak leži u tome što je nedovoljno pažnje posvećivano jezičkim odlikama inputa u knjigama za decu ranog i predškolskog uzrasta. Prevazilaženje ovog nedostatka iziskuje sačinjavanje dovoljno velike i reprezentativne kolekcije mašinski čitljivih tekstova, tj. korpusa, nad kojim će sistematske lingvističke analize biti sprovođene.

Upravo sa tim ciljem, u okviru Katedre za opštu lingvistiku i Centra za primenjenu lingvistiku Filološkog fakulteta Univerziteta u Beogradu, započet je projekat formiranja Korpusa knjiga za decu ranog i predškolskog uzrasta - *DeciKo* 2022. godine.

Korpus *DeciKo* osmišljen je kao dinamički specijalizovani korpus, koji se dopunjava na godišnjem nivou. Trenutno je strukturiran na osnovu tri vanjezička kriterijuma: (1) žanra, (2) jezika originala i (3) veka prvog izdanja uzoraka. Dostupan je na upit autorima i nalazi se u TEI XML formatu. Trenutno sadrži preko 250 knjiga za decu. Knjige su putem digitalnih alata prevođene u tekstualni format, a zatim su greške ručno ispravljane.

U ovom izlaganju predstavićemo trenutno stanje korpusa *DeciKo*, i opisati probleme i izazove na koje nailazimo tokom konstruisanja ovog korpusa.

Glavni problem u konstruisanju korpusa *DeciKo* predstavlja nemogućnost da se uzorci za korpus izaberu/podele prema uzrastu kome su namenjeni. Različiti izdavači na različite načine razvrstavaju (čak i iste) knjige po uzrastima, dok rezultati sprovedenih pilot istraživanja jezičkih karakteristika uzoraka pokazuju značajna odstupanja od ustanovljenog toka usvajanja prvog jezika. Da ovo nije lokalni problem pokazuje i to što svi postojeći korpusi knjiga za decu ranog i predškolskog uzrasta sadrže samo žanrovsku ali ne i uzrasnu podelu uzoraka u svom sastavu. Budući da je *DeciKo* osmišljen kao korpus sa punim potencijalom za istraživanja u oblasti usvajanja prvog jezika i usvajanja pismenosti, neophodno je naći način za prevazilaženje ovog problema.

Ključne reči: *DeciKo, korpus dečjih knjiga, problem uzorkovanja, usvajanje prvog jezika*

Literatura

[1] Nation, K., Dawson, N., & Hsiao, Y. (2022). Book Language and Its Implications for Children's Language, Literacy, and Development. *Current Directions in Psychological Science, 31*, 375–380.

Dušanka Vujović
Hankuk University of Foreign Studies in Seoul
Faculty of East European & Balkan Studies, Department of South Slavic Studies
E-mail: dusanka.vujovic@ff.uns.ac.rs

Branko Milosavljević
Univerzitet u Novom Sadu - Fakultet tehničkih nauka
E-mail: mbranko@uns.ac.rs

Korpus Srpko kao tehnološka osnova za izradu rečnika savremenog srpskog jezika

Digitalizacija jezičkih resursa omogućava lak i brz pristup jezičkim podacima olakšavajući tako lingvistička istraživanja kao i analizu jezičkih fenomena. Ona je, takođe, od ključnog značaja za savremene leksikografske projekte jer obezbeđuje obilje pažljivo odabrane leksičke građe, lakšu pretragu i ekscerpciju primera kao i njihovu ugradnju u rečnik. Izgradnja elektronskog jezičkog korpusa Srpko predstavlja temelj za izradu novog Rečnika savremenog srpskog jezika Matice srpske. Elektronski jezički korpus omogućava jednostavno i brzo pretraživanje tekstualnih podataka i njihovu inkorporaciju u sadržaj rečničkog teksta. U radu se govori o principima kao i o idejnim i tehničkim rešenjima u vezi sa organizacijom korpusa,

prikazuje se njegova struktura koja se sastoji od književnog, novinskog, naučnog, administrativnog i razgovornog potkorpusa i način prikupljanja podataka i formiranja baze podataka za svaki od njih budući da se radi o podacima dostupnim u različitim formatima kao što su pisani digitalizovani i nedigitalizovani tekstovi, audio i video zapisi. U cilju brže i lakše pretrage korpusa, ali i njegove kontrole u program su ugrađene različite funkcionalne pretrage i zahtevi za filtriranje rezultata što je omogućilo relativno smanjivanje šuma, odnosno redundantnih informacija koje zagušuju i otežavaju pretrage. U program je ugrađena i funkcija dobijanja različitih izveštaja na osnovu zadatih filtera. To mogu biti već pripremljeni izveštaji dobijeni na osnovu unapred određenih parametara ili dinamički izveštaji koji mogu da se kreiraju po potrebi na osnovu zadatih filtera. Ovako koncipiran korpus pruža osnovu za sistematsko organizovanje i predstavljanje leksičke baze podataka u cilju izrade rečnika.

Ključne reči: *srpski jezik, korpus srpskog jezika, digitalizacija, rečnik*

Kristina Ilić
Univerzitet u Beogradu, Filološki fakultet
E-mail: kristina.ilic997@gmail.com

Značaj paralelnih korpusa za istraživanje frazemskih konstrukcija u nemačkom i srpskom jeziku

Paralelni korpusi su od velikog značaja za konstrastivna istraživanja frazemskih konstrukcija u različitim jezicima. Prema Dobrovoljskom (npr. 2011, str. 114), frazemske konstrukcije definisane su kao leksički delimično specifični parovi oblika i značenja (= konstrukcije), čija semantika proizlazi ne samo iz leksičkog značenja njihovih komponenti i načina njihove sintaktičke povezanosti, već i iz same konstrukcije kao celine. Frazemske konstrukcije su idiomatske kombinacije reči koje se sastoje od određenih leksičkih elemenata, tzv. sidra i određenog broja praznih mesta, tzv. slotova. Ove praznine popunjavaju se tzv. filerima, pri upotrebi konstrukcije. (Dobrovol'skij, 2011, str. 114). Kako potvrđuje i Đurčo, savremeni jednojezični i višejezični korpusi pružaju nove alate za poređenje podataka iz velikih korpusa dva jezika, koji se sastoje od nezavisnih tekstova (Đurčo, 2018, str. 117). Konstrukciona gramatika je mlada grana lingvistike, a istraživanje frazemskih konstrukcija uz pomoć korpusa, a naročito paralelnih korpusa u slučaju kontrastivnih istraživanja, se sve više razvija. Ovo je naročito slučaj ako je cilj pretrage u korpusu dolaženje do prevodnih ekvivalenata konstrukcija na izvornom jeziku u paralelnim tekstovima na ciljnom jeziku, pri čemu su ti prevodni ekvivalenti u ciljnom jeziku često i sami konstrukcije, što doprinosi istraživanju frazemskih konstrukcija i u ciljnom jeziku. U ovom trenutku može se reći da su paralelni korpusi nemačkog i srpskog jezika slabije zastupljeni, za razliku od paralelnih korpusa nemačkog i drugih većih svetskih jezika, kao što je COMBIDIGILEX ili EuReCo, ili paralelnih korpusa drugih većih svetskih jezika i srpskog jezika, kao što je SrpEngKor, SrpFranKor, Evroteka ili ParCoLab. Ovo izlaganje ima za cilj da predstavi dosada dostupne resurse u vidu paralelnih korpusa nemačkog i srpskog jezika, kao što su SrpNemKor u okviru alata Bibliša, korpusi koji su deo Sketch Engine-a ili InterCorp u okviru češkog Nacionalnog korpusa. Osim toga, pokazaće moguće načine njihove upotrebe u kontrastivnom istraživanju frazemskih konstrukcija na konkretnim primerima i ukazaće na to za čime još postoji potreba i predložiće izgled, obim, mogućnosti pretrage itd. optimalnog resursa za takva istraživanja. Ovo istraživanje sprovodi se u okviru COST akcije PhraConRep Višejezični repozitorijum frazemskih konstrukcija na jezicima centralne i istočne Evrope ("A Multilingual Repository of

Phraseme Constructions in Central and Eastern European Languages"), koja ima za cilj između ostalog i da kontrastivno istraži frazemske konstrukcije nemačkog i srpskog jezika, pri čemu upotreba različitih alata i paralelnih korpusa, predstavljenih u ovom istraživanju, može odigrati presudnu ulogu u dolaženju do tog cilja.

Ključne reči: *Konstrukciona gramatika, korpusna lingvistika, frazemske konstrukcije, srpski jezik, nemački jezik*

Literatura

[1] Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch. & A. Ziem (eds.), *Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze* (pp. 110–130). Tübingen: Stauffenburg.

[2] Ďurčo, P. (2018). Vom Nutzen der vergleichbaren Korpora bei der kontrastiven lexikographischen Erfassung von Mehrworteinheiten. In V. Jesenšek & M. Enčeva, (eds.), *Wörterbuchstrukturen zwischen Theorie und Praxis* (pp. 107–118). Berlin, Boston: De Gruyter.

Nevena Ceković
Univerzitet u Beogradu – Filološki fakultet
E-mail: n.cekovic@fil.bg.ac.rs

Revizija grešaka u učeničkom ITALSERB korpusu

Ortografska transkripcija predstavlja krucijelan korak u kreiranju upotrebljivog govornog korpusa. Složenost realizacije takvog metodološkog postupka proizvodi, međutim, sasvim očekivano, mogućnosti za niz potencijalnih grešaka. Tim pre ukoliko je učenički ili korpus drugog jezika u pitanju, a naročito kada izrada jednog takvog korpusa predstavlja pionirski poduhvat na polju primenjene korpusne lingvistike u Srbiji. Homogenost transkripcionih zapisa spram činjenice da u tom postupku nužno učestvuje više istraživača, takođe predstavlja poseban metodološki izazov.

U radu se pruža uvid u proces revizije ITALSERB (ITALiano dei SERBofoni) korpusa srbofonih studenata italijanskog kao stranog jezika koji se od 2010. godine realizuje na Katedri za italijanistiku Filološkog fakulteta Univerziteta u Beogradu. Ovaj korpus velikih dimenzija sačinjen je od oko 25 sati audio snimaka asimetričnih interakcija u kojima je učestvovalo preko 170 ispitanika, različitih nivoa jezičkih kompetencija (od A2 do C1). Korpus se trenutno nalazi u fazi revizije transkripcionog procesa, kao i morfosintaksičke anotacije prikupljene jezičke građe.

U radu se posebno razmatraju korekcije grešaka nastalih kao posledica transkripcionog procesa za razliku od onih, karakterističnih za učeničke korpuse, koje oslikavaju tipične stadijume u razvoju kompetencije i koje stoga pružaju dragocen uvid u specifične odlike međujezika. Cilj rada jeste da kroz oslikavanje pojedinih primera transkripata ekscerpiranih iz korpusa za potrebe revizije pruži smernice za dobru praksu u konstituisanju (transkripciji i anotaciji) korpusa drugog jezika.

Ključne reči: *ITALSERB, učenički korpus, korpus drugog jezika, italijanski kao strani jezik, srbofoni studenti, transkripcija, revizija*

Јелена Марковић
Универзитет у Источном Сарајеву
E-mail: jelena.markovic@ff.ues.rs.ba

Ученички корпуси србофоних говорника енглеског језика у контрастивној анализи међујезика

Развој контрастивне анализе међујезика (енг. Contrastive Interlanguage Analysis), која је настала као резултат потребе иновирања контрастивне методологије у оквиру примењене лингвистике, у директној је зависности од појаве и развоја ученичких корпуса, попут међународног ученичког корпуса ICLE (The International Corpus of Learner English). Данас је контрастивна анализа међујезика афирмисана као незаобилазна методологија у дисциплини истраживања ученичких корпуса (енг. Learner Corpus Research).

И на српском говорном подручју остварен је напредак у развоју контрастивне анализе међујезика и ученичких корпуса. У том правцу велики значај има формирање корпуса србофоних говорника ICLE-SE (International Corpus of Learner English – Serbian), који је заправо један од двадесет и пет поткорпуса поменутог корпуса ICLE, одређених матерњим језиком говорника. Осим њега, важно је истаћи и формирање комплеменарног корпуса мањег обима, КорСАнг (Корпус студената англистике), који укључује преводе студената англистике у оба смера, као и аргументативне саставе студената писане на српском језику као матерњем. Два поменута србофона корпуса описујемо и из углова метаподатака које поседују и актуелног степена анотације.

Након увода, пружамо увид у резултате досадашњих истраживања у оквиру контрастивне анализе међујезика на српском говорном подручју које укључују поменуте србофоне ученичке корпусе, у циљу критичког осврта и дискусије могућих праваца развоја. Говоримо, између осталог, о истраживањима метадискурсних обележја, дискурсних конектора и феномену лексичке неодређености, уз анализу фреквенцијских листа речи у поменутим србофоним и референтним изворним корпусима. Излагање завршавамо истицањем недовољно коришћеног научног потенцијала постојећих србофоних корпуса, уз предлоге за њихову популаризацију, као и могућа вишеструка проширења

Кључне речи: *ученички корпус, ICLE-SE, КорСАнг, контрастивна анализа међујезика, метаподаци, варијабле, дискурс*

Литература

[1] Марковић, Ј. (2017). Лични метадискурс у писању код неизворних и изворних говорника енглеског језика. *Филолог, VII* (15), 44–60.
[2] Марковић, Ј. (2018). Употребе глагола *make* у писању на енглеском језику као страном код изворних говорника српског језика (корпуснолингвистичка анализа). *Зборник Матице српске за филологију и лигвистику, LXI* (1), 165–180.
[3] Марковић, Ј. (2019). *Кроз призму контрастивне анализе међујезика.* Пале: Филозофски факултет.
[4] Марковић, Ј. (2020). Концесивни конектори *though* и *however* у писању на енглеском језику код изворних и неизворних говорника. *Филолог*, 21, 13–35.
[5] Марковић, Ј. и Станковић, Р. (2021). *Ја/ти/ми/ви* у дискурсној компетенцији у светлу контрастивне анализе међујезика. *Методички видици*, 12, 95–119.
[6] Радоња, М. и Шућур, С. (2021). О Корпусу студената англистике (КорСАнг) и могућностима његове софтверске експлоатације. *Infotheca – Journal for Digital*

Humanities, 21(1), 37–58.Ädel, A. (2006). *Metadiscourse in L1 and L2 English* (Studies in Corpus Linguistics ed., Vol. 24). Amsterdam/Philadelphia: John Benjamins.
[7] Deshors, S. C., Götz, S., & Laporte, S. (2016). Linguistic innovations in EFL ansd ESL: Rethinking the linguistic creativity of non-native English speakers. *International Journal of Learner Corpus Research, 2*(2), 131–150.
[8] Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. In G. Guilquin, S. Papp, & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 3–33). Amsterdam and New York: Rodopi.
[9] Granger, S. (2012). How to use second and foreign language learner corpora. In A. Mackey, & S. M. Gass (Eds.), *Research methods in Second Language Acquisition: A Practical Guide* (pp. 7–29). Malden: Blackwell.
[10] Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.
[11] Lee, D. Y., & Chen, S. X. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing, 18*, 281–296.
[12] Marković, J. (2019). "It is a thing that gives you…": The lexeme thing(s) as the Serbian EFL 'teddy bear'. *Komunikacija i kultura online, 10,* 19–37.
[13] Marković, J. (2021). *I, you,* and *we* in Serbian EFL Argumentative Writing from the Essay Title Perspective. *Folia Linguistica et Litteraria, 36,* 271–290. doi: 10.31902/fll.36.2021.16
[14] Nesselhauf, N. (2005). *Collocations in a Learner Corpus* (Studies in Corpus Linguistics ed., Vol. 14). Amsterdam/Philadelphia: John Benjamins.
[15] Šućur, S. (2019). Distribucija frazalnih glagola u pisanju na engleskom kao stranom kod srbofonih govornika. *Komunikacija i kultura online, 10,* 120–143.
[16] Tomović, N., & Marković, J. (2020). The Status of English in Serbia. In S. Granger, M. Dupont, F. Meunier, H. Naets, & M. Paquot (Eds.), *The International Corpus of Learner English. Version 3.* (236–242). Louvain-la-Neuve: Presses universitaires de Louvain.

Saša Moderc
Univerzitet u Beogradu – Filološki fakultet
E-mail: moderc.sasa@gmail.com

Italijanski klitici i tagovanje: korisnička iskustva u radu s korpusom Serbitacor3 Corpus

Italijanski jezik poseduje jedanaest klitika u raznim službama: lične zamenice, priloške odredbe, markera za pasiv, markera za bezličnu strukturu; klitici se upotrebljavaju i pleonastično i sastavni su deo prokomplementarnih glagola (npr. farcela 'uspeti', od osnovnog glagola fare 'činiti'). Mogu da stoje u enklitičkom, proklitičkom i mezoklitičkom položaju u odnosu na glagol; mogu da se spajaju u grupe od dva ili tri klitika i kao grupa mogu zauzimaju jedan od navedenih položaja, ali sama grupa može i da se rastavi, tako da jedan klitik zauzima proklitički položaj a drugi enklitički. Većina klitika je polifunkcionalno: zaključivanje o njihovoj službi zasniva se na sintaksičkim, semantičkim i pragmatičkim faktorima i zahteva naprednu lingvističku kompetenciju. Takvu kompetenciju poseduju prevodioci književnih dela na osnovu čijih je tekstova sastavljen korpus srpsko-italijanskih paralelnih tekstova

SerbItaCor3. Na primer, niz klitika lo si è visto se javlja tri puta u korpusu i u sva tri slučaja je preveden ispravno na srpski. Klitik lo je u jednom slučaju u službi lične zamenice (su ga videli), u dva u službi pokazne zamenice (kao što smo Ø videli, već smo to videli). Međutim, u sva tri slučaja u korpusu SerbItaCor3 lo je tagovan kao lična zamenica, a tagovanje je obavljeno upotrebom TreeTagger-a autora Ahima Štajna. U radu s navedenim korpusom primećene su i druge nepreciznosti u tagovanju klitika, što je naročito osetljivo kada je reč o klitiku si, koji je na nekim mestima netačno označen kao lična zamenica (u nizu si guarda, gde si može da bude ili povratna zamenica, ili marker za pasiv ili marker za bezličnu strukturu); data mesta su prevedena na srpski precizno. Ove i druge nepreciznosti u tagovanju klitika umanjuju pouzdanost lingvističkih podataka kojima je opremljen korpus jer se time smanjuje domet korpusa u domenu glotodidaktike. Ipak, zahvaljujući pažljivim prevodima moguće je uočiti pravu funkciju klitika i s tim podatkom doprineti unapređenju tagera kroz dopunjavanje uputstvima u samom TreeTaggeru (u pogledu valentnosti glagola, argumenata i odredbi koje mogu da stoje uz njega). U radu su predstavljeni primeri nepreciznog tagovanja klitika i korektivnih prevodilačkih rešenja; prikazuju se lingvistički parametri pomoću kojih je moguće izvesti instrukcije za unapređenje postojećih tagera ili za pisanje novog tagera za italijanski jezik.

Ključne reči: *klitici, italijanski jezik, srpski jezik, korpus, tagovanje, unapređenje tagera*

Literatura

[1] Bentley, D. (2006). Split Intransitivity in Italian. Berlin-New York. Mouton de Gruyter.
[2] Dell'Orletta F. (2009). Ensemble system for Part-of-Speech tagging. In: Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian. Reggio Emilia, Italy.
[3] Moderc S., Stanković R., Tomašević A., Škorić M. (2023). An italian-serbian sentence aligned parallel literary corpus. Review of the National Center for Digitization, 43.
[4] Moderc, S. (2021). I clitici italiani. Usi, ambiguità, interpretazioni. Volume primo: il sistema dei clitici. Beograd. Filološki fakultet.
[5] Moderc, S. (2021). I clitici italiani. Usi, ambiguità, interpretazioni. Volume secondo: i nessi di clitici. Beograd. Filološki fakultet.
[6] Moderc, S. (2015). Gramatika italijanskog jezika. Morfologija s elementima sintakse. Beograd. Luna crescens.
[7] Renzi, L. (1988). Grande grammatica italiana di consultazione. Vol. 1. Bologna. Il Mulino.
[8] Russi C. (2008). Italian Clitics. An Empirical Study. Berlin-New York. Mouton de Gruyter.
[9] Salvi, G., Vanelli, L. (2004). Nuova grammatica italiana. Bologna. Il Mulino.
[10] Schmid, H. (2013). *Probabilistic part-of speech tagging using decision trees*, In New methods in language processing, 5.
[11] Serianni, L. (1989). Grammatica italiana. Italiano comune e lingua letteraria. Torino. Utet.
[12] Tamburini F. (2009). PoS-tagging Italian texts with CORISTagger. In: EVALITA 2009. Workshop on Evaluation of NLP and Speech Tools for Italian. vol. 1. Reggio Emilia, Italy, December 2009.
[13] Tamburini, F. (2000). Annotazione grammaticale e lemmatizzazione di corpora in italiano. In: R. Rossini, Linguistica e informatica: multimedialità, corpora e percorsi di apprendimento. Bulzoni, Roma, 2000.

Olja Perišić
Univerzitet u Torinu
E-mail: olja.perisic@unito.it

Korpusi za učenje srpskog jezika kao stranog u eri veštačke inteligencije

Uvođenje korpusa u nastavu stranih jezika vezano je za devedesete godine prošlog veka i pristup poznat kao DDL (Data Driven Learning) u didaktici engleskog kao stranog jezika. Dok istraživači aktivni u ovoj oblasti preispituju budućnost korpusnih alata u eri veštačke inteligencije (Crosthwaite i Baisa 2023, Flowerdew 2024), srpski jezik ostaje u povoju, barem što se tiče praktične upotrebe korpusa u nastavi. Naime, uočava se značajan jaz između, sa jedne strane, istraživačkog delovanja stručnjaka za informatiku (posebno onih okupljenih oko udruženja JeRTeh) koji prate svetske trendove i aktivno rade na razvijanju korpusnih alata i, sa druge strane, nastavnog osoblja koje bi te alate moglo da koristi u radu sa učenicima. Pored neophodnog afiniteta za ovu vrstu nastave, jedan od uzroka ovom raskoraku je početno ulaganje u savladavanje veštine pretrage korpusima, kao i usvajanje teorijskog i metodološkog pristupa koji omogućava rad u učionici. Iskustvo je pokazalo da studenti srpskog kao stranog jezika relativno brzo savladavaju veštinu korišćenja korpusa i da su uglavnom kreativni u samostalnim istraživanjima (Perišić 2023). Pokazalo se takođe da zahvaljujući korpusima studenti od početnog nivoa razvijaju osećaj za jezik, sklonost ka istraživačkom radu i motivaciju za učenje srpskog. U eri ubrzanog razvoja veštačke inteligencije i rastućeg trenda opadanja broja studenata filoloških studija, ne samo za naš jezik, u radu ćemo istaći prednosti korpusnih istraživanja u učenju stranih jezika, sa naglaskom na srpski kao strani i ukazati na izazove ali i prednosti korpusnih alata u odnosu na generativne jezičke modele.

Ključne reči: *korpusi, srpski kao strani, Data Driven Learning, JeRTeh, veštačka inteligencija*

Literatura

[1] Crosthwaite Peter, Baisa Vit (2023). *Generative AI and the end of corpus-assisted data-driven learning? Not so fast!*, "Applied Corpus Linguistics 3", https://doi.org/10.1016/j.acorp.2023.100066
[2] Flowerdew John (2024), Data-driven learning: From Collins Cobuild Dictionary to ChatGPT, *Language Teaching*, 1–18.
[3] Perišić Olja (2023), *Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico*, Alessandria: Edizioni dell'Orso.

Simon Krek
"Jožef Stefan" Institute, Artificial Intelligence Laboratory
University of Ljubljana, Centre for Language Resources and Technologies
E-mail: simon.krek@ijs.si

Carole Tiberius
Faculty of Humanities, Leiden University
Centre for Linguistics
E-mail: carole.tiberius@ivdnt.org

Jaka Čibej
Faculty of Arts, University of Ljubljana
Centre for Language Resources and Technologies, University of Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si

Ana Ostroški Anić
Department of General Linguistics
Institute for the Croatian Language
E-mail: aostrosk@ihjj.hr

Ranka Stanković
University of Belgrade, Faculty of Mining and Geology
Chair for Mathematics and Informatics
E-mail: ranka@rgf.rs

Proširenje paralelnog semantički anotiranog korpusa ELEXIS-WSD na južnoslovenske jezike: izazovi, rezultati i planovi

ELEXIS-WSD je paralelni semantički anotiran otvoreni korpus (Martelli et al. 2021; Martelli et al. 2023) razvijen u okviru ELEXIS projekta koji u verziji 1.1 sadrži 2024 rečenice za svaki od 10 jezika: bugarski, danski, engleski, španski, estonski, mađarski, italijanski, holandski, portugalski i slovenački. Unutar rečenica, svakoj reči koja nosi značenje (imenica, pridev, glagol i prilog) dodeljeno je odgovarajuće značenje iz jednog od 10 repozitorijuma značenja koji sadrže definicije.

U kontekstu zadatka 2.2 COST akcije UniDive (CA21167), korpus će biti proširen sa nekoliko novih jezika, uključujući najmanje dva južnoslovenska jezika: hrvatski i srpski. Proces proširenja korpusa uključuje nekoliko različitih faza od prevođenja do tokenizacije, lematizacije i obeležavanja vrste reči, potom obeležavanja imenovanih entiteta i polileksemskih izraza, i konačno, pridruživanja značenja reči. U ovom radu predstavljamo izazove sa kojima se susrećemo u ovim različitim fazama pri proširenju korpusa srpskim i hrvatskim jezikom, kao i u procesu dodavanja novih slojeva anotacija za slovenački deo korpusa.

Srpski dodatak višejezičnom anotiranom korpusu ELEXIS započeo je automatskim prevođenjem rečenica, nakon čega je usledilo naknadno ručno prilagođavanje. Tokenizacija, obeležavanje vrste reči, lematizacija, prepoznavanje i povezivanje imenovanih entiteta (NE) su takođe ručno proveravani (Krstev et al. 2024). Obležavanje obuhvata takođe i prepoznavanje višečlanih izraza (MWE). Biće reči o prvim koracima i izazovima u izgradnji srpskog repozitorijuma značenja, a biće analizirani i neki rezultati koji se odnose na MWE i NE. Kada

bude završen, ELEXIS-WSD-SR korpus će biti prvi korpus sa anotacijama zasnovanim na srpskog Vordnetu (SrpWN).

Hrvatski dodatak korpusa ELEXIS započeo je istom procedurom automatskog prevođenja i ručne validacije rečenica. Međutim, bio je potreban dodatni rad na pripremi repozitorijuma značenja za hrvatski jezik koji će se koristiti za razlučivanje značenja reči. Resurs koji se koristi kao osnova za zadatak je rečnik u XML formatu, čije kategorije podataka nisu strukturisane u skladu sa postojećim leksičkim ili leksikografskim modelima podataka. Zbog toga su podaci prvo raščlanjeni u zasebne kategorije podataka, nakon čega je usledila opsežna provera da su različite oznake (npr. gramatičke i oznake upotrebe) prepoznate prema njihovoj prvobitnoj nameni. Takođe će biti reči o proširenju resursa i njegovoj konverziji u otvoreni repozitorijum.

Kao i prethodna verzija, prošireni korpus i odgovarajući repozitorijumi značenja biće dostupni u CLARIN.SI repozitorijumu pod CC-BI-SA 4.0.

Zahvalnica: *Ovo istraživanje je podržala COST akcija CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive). Autori se zahvaljuju za finansijsku podršku Slovenačke agencije za istraživanje i inovacije (osnovno finansiranje istraživanja br. P6-0411 – Jezički resursi i tehnologije za slovenački jezik).*

Ključne reči: *semantička anotacija, paralelni korpus, značenja, južnoslovenski jezici, slovenački, hrvatski, srpski*

Literatura

[1] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Veronika Lipp, Tamás Váradi, András Győrffy, and László Simon. "Designing the ELEXIS parallel sense-annotated dataset in 10 European languages." (2021): 377-395.

[2] Cvetana Krstev, Ranka Stanković, Aleksandra Marković, Teodora Mihajlov “Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities, Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024), pp. 74-84, SIGLEX, ACL, UniDive CA21167. eds.: Bhatia, Archna et al. virtual presentation, May 25, 2024. https://aclanthology.org/2024.mwe-1.15.pdf

[3] Martelli, Federico, Roberto Navigli, Simon Krek, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langemets, Kristina Koppel, Tiiu Üksik, Kaja Dobrovoljc, Rafel Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamás Váradi, András Győrffy, László Simon, Valeria Quochi, Monica Monachini, Francesca Frontini, Carole Tiberius, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, Tina Munda, Iztok Kosem, Rebeka Roblek, Urška Kamenšek, Petra Zaranšek, Karolina Zgaga, Primož Ponikvar, Luka Terčon, Jonas Jensen, Ida Flörke, Henrik Lorentzen, Thomas Troelsgård, Diana Blagoeva, Dimitar Hristov, Sia Kolkovska, 2023: Parallel sense-annotated corpus ELEXIS-WSD 1.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1842.

Душко Витас, Ранка Станковић, Цветана Крстев
Друштво за језичке реурсе и технологије JERTEX
E-mail: {vitas|cvetana|ranka}@jerteh.rs

Многа лица СрпКор-а

Акроним СрпКор означава фамилију електронских корпуса савременог српског језика чија је изградња почела крајем седамдесетих година прошлога века, а која је постала шире видљива заинтересованој истраживачкој заједници објављивањем његове прве верзије на вебу 2002. године. У овом дугом периоду, посебно пре појаве корисних текстуелних ресурса на вебу, развој корпуса се састојао у прикупљању и обради грађе као и у развоју метода обраде корпуса. Наиме, електронски корпус није само колекција текстова у дигиталном облику (како се то, на пример, наводи у (Добрић 2012)), већ подразумева више компонената које ће заједно овакву колекцију учинити корисном у језичким и другим истраживањима. Ове компоненте, поред самих текстова, чине, пре свега, софтверска подршка организацији и експлоатацији колекције текстова и средства за различите нивое анотације текстова који ће се наћи у корпусу (Витас 2023).

СрпКор је, водећи рачуна о овим компонентама, прошао различите метаморфозе током своје изградње које пружају слику како о еволуцији софтверске подршке за конструкцију и експлоатацију корпуса, као и о развоју система анотација на различитим нивоима (мета-подаци, морфолошко обележавање, лематизација, именовани ентитети, итд).

Крајње скромни услови (у поређењу са другим срединама, како у броју истраживача укључених у изградњу корпуса, додељеним финансијским средствима из различитих извора, расположивом опремом) су наметнули стратегију поступног развоја корпуса која је подразумевала да ће се нове верзије корпуса ослањати на материјал припремљен и употребљен у оним верзијама које су јој претходиле.

У раду ће бити илустрована еволуција у развоју СрпКор-а почев од његове прве верзије до данас пратећи упливе различитих средстава која су коришћена у изградњи појединачних верзија, као и промене димензија и система анотације текстова. Посебно ће бити описана структура појединачних верзија корпуса, њихове димензије, обухваћени временски период и ниво анотације.

Основне замисли приликом конципирања корпуса су прво изложене у (Витас, Поповић 2023), а затим у (Утвић 2013) где су описани бројни детаљи за верзију СрКор-а из 2013. године. Интеракције корпуса са речницима су разматране у (Krstev, Vitas 2005), (Vitas, Krstev 2012).

Значајно је напоменути да су у СрпКор унети и текстови на српском језику из паралелизованих корпуса који су настајали упоредо са СрпКор-ом. На овај начин је делимично компензован утицај веб-садржаја на састав корпуса. С друге стране, овакви текстови који су, по правилу, изузетно значајни у културном смислу, јер не само да нису присутни у грађи са веба, већ обично не улазе ни у традиционалне лексикографске корпусе. Њих чине одабрани научни, књижевни, филозофски, антрополошки, историјски и слични текстови преузети из угледних едиција.

Даљи рад на развоју овог корпуса обухватиће са једне стране обогаћивање метаподатака, допуну анотација и унос нових садржаја. Обогаћивање метаподатака омогућиће креирање подкорпуса по различитим димензијама: по изговору, периоду,

домену, уз до сада расположиве по аутору, регистру, годинама. Уз поделу на реченице и допуну анотација именованим ентитетима, у плану је обогаћивање и граматичким информацијама. Унос нових садржаја проширује временску димензију припремом романа, путописа, мемоара, историјских новина које представљају драгоцени материјал не само из филолошког већ и из културно-историјског аспекта, уз уобичајено допуњавање (изабраним) садржајима са веба.

Спрега лексичке базе Лексимирка и фамилије корпуса СрпКор је двосмерна (Lazić, Škorić 2020). Кроз интерфејс Лексимирка је могућ директан увид у примере употребе речи у контексту или у синтаксичким обрасцима. Систем за лематизацију се унапређује из верзију у верзију, у чему специјалну улогу имају електронски морфолошки речници српског језика који коришћењем система Unitex обезбеђују генерисање свих флективних облика лема.

Кључне речи: *СрпКор, корпуси, српски, лематизација, Лексимирка*

Захвалница: *Ово истраживање је подржао Фонд за науку Републике Србије, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Литература

[1] Cvetana Krstev, Duško Vitas, “Corpus and Lexicon - Mutual Incompletness”, in Proceedings of the Corpus Linguistics Conference, 14-17 July 2005, Birmingham, eds. Pernilla Danielsson and Martijn Wagenmakers, ISSN 1747-9398, http://www.corpus.bham.ac.uk/PCLC/, 2005.

[2] Dobrić, Nikola. "Savremeni jezički korpusi na Zapadnom Balkanu–Istorijat, trenutno stanje i budućnost (Language Corpora in the West Balkans–History, Current State and Future Perspective)." Slavistična revija 60 (2012): 677-692.

[3] Душко Витас, Љубомир Поповић, „Конспект за изградњу референтног корпуса српског стандардног језика“, Научни састанак слависта у Вукове дане 31/1 - МСЦ, Београд, 2003, стр. 221 - 227.

[4] Душко Витас, Белешке о ручној и аутоматској обради српског језика, Језик Данас, бр. 22, 2023, Матица Српска, Нови Сад.

[5] Duško Vitas, Cvetana Krstev “Tvorbeni obrasci u elektronskom rečniku srpskog jezika”, Međunarodni komitet slavista. Komisija za tvorbu reči. Međunarodna naučna konferencija Tvorba reči i njeni resursi u slovenskim jezicima (14), pp. 515-525, 2012, Filološki fakultet Univerziteta u Beogradu, Beograd, ISBN 978-86-6153-116-3

[6] Utvić, Miloš. 2013. “Izgradnja referentnog korpusa savremenog srpskog jezika.” PhD diss., Univerzitet u Beogradu, Filoloˇski fakulte. https://nardus.mpn.gov.rs/handle/123456789/4091

[7] Biljana Lazić, Mihailo Škorić. "From DELA Based Dictionary to Leximirka Lexical Database" in Infotheca, Faculty of Philology, University of Belgrade (2020). https://doi.org/10.18485/infotheca.2019.19.2.4

Саша Марјановић
Универзитет у Београду – Филолошки факултет
E-mail: sasa.marjanovic@fil.bg.ac.rs

Дејан Стошић
Универзитет Тулуз – Жан Жорес и CLLE (UMR 5263 CNRS)
E-mail: dejan.stosic@univ-tlse2.fr

Збирке текстова на српскоме језику у деветојезичном паралелном корпусу *ParCoLab*

Вишејезични — деветојезични — паралелни корпус ParCoLab (Stosic i dr. 2015) резултат је билатералне сарадње истраживача с Универзитета Тулуз – Жан Жорес у Тулузу и Катедре за романистику Филолошког факултета Универзитета у Београду, око којих се одскора окупљају и сарадници из других научноистраживачких организација у Француској (Универзитет у Поатјеу, Универзитет у Стразбуру и Универзитет у Корзици). Корпус се похрањује у оквиру француске дигиталне инфраструктуре HumaNum, а слободно — и без израде налога — може се користити путем сучеља посебно израђеног за потребе претраживања корпуса на интернетској страници: http://parcolab.univ-tlse2.fr.

С развијањем корпуса започело се 2007. године упаривањем књижевних текстова на српскоме и француском језику, а од 2010. у корпус се укључују и текстови на енглеском језику (Miletić i dr. 2014; Miletić i dr. 2017; Marjanović i dr. 2018; Stosic i dr. 2019; Марјановић и др. 2019). Додавање нових језика настављено је 2018. године, кад су у корпус убачени текстови на окситанском језику, 2020. кад је започет унос шпанскога, 2022. и 2023. кад су укључени с једне стране документи на поатвенском, а с друге на алзашкоме и корзиканском језику, који су, уз окситански, неки од регионалних језика што чине језичко наслеђе Француске (Stosic i dr. 2024); коначно, 2024. године почело се с развојем италијанског поткорпуса интегрисањем обосмерно упарених текстова корпуса SerbItaCor3 (Moderc i dr. 2023). Поред тога, током целокупног периода радило се и на ширењу текстних домена и жанрова (нпр. Terzić i dr. 2020), па су тако, поред текстова лепе књижевности, у корпус унети и вишејезично упарени транскрипти целовечерњих и анимираних филмова, новински текстови, текстови вишејезичких интернетских портала, преведени религијски, биолошки, кинематографски текстови, текстови спортских регулатива и других правних прописа, политички говори и друго. Корпусни су текстови разврстани, у зависности од домена и порекла података, у одговарајуће посебне збирке. Збирке тренутно упарених текстова и текстова доступних за претраживање на свим обухваћеним језицима досежу 50.000.000 речи. У овоме излагању представљају се управо збирке текстова вишејезичнога паралелног корпуса ParCoLab у којима је српски језик како изворни, тако и циљни, затим се указује на њихов обим заступљености, те се закључује могућностима примене у неколико филолошких подручја.

Кључне речи: *ParCoLab, паралелни корпус, српски језик, француски језик, енглески језик, шпански језик, италијански језик, регионални језици у Француској*

Ресурси за обраду јужнословенских језика

Jelena Lazarević
Univerzitet u Beogradu, Filološki fakultet, doktorand
E-mail: jelazarevic1@gmail.com

Olivera Kitanović
Univerzitet u Beogradu, Rudarsko-geološki fakultet
E-mail: olivera.kitanovic@rgf.bg.ac.rs

Kontrastivna analiza sintaksičkih obrazaca u komparabilnim korpusima fudbala na španskom i srpskom jeziku

Cilj rada je istraživanje kolokabilnosti kao načina na koji se leksičke jedinice povezuju sa rečima iz različitih kategorija, formirajući veće jedinice. Istraživanje semantičkih i sintaksičkih principa ovih kombinacija u španskom i srpskom jeziku fudbala izvedeno je na komparabilnim fudbalskim korpusima SrFudKo i EsFudko, razvijenim u okviru doktorske disertacije Jelene Lazarević pod nazivom: Jezičke odlike diskursa novih medija o fudbalu: kontrastivna analiza na korpusu srpskog i španskog jezika.

Korpus fudbala SrFudKo, kreiran na osnovu tekstova o fudbalu sa pet srpskih veb-portala: B92, Blic, Mondo, Politika i Sport klub, sadrži 10.100.553 tokena, od toga 8.618.426 reči. Korpus EsFudKo o fudbalu na španskom jeziku potiče od tekstova sa dva španska veb-portala: Marca fútbol i Mundo deportivo, a sadrži 9.106.812 tokena, od čega 8.024.164 reči. Oba korpusa nad kojima su primenjene metode korpusne lingvistike za ekstrakciju podataka se nalaze platformi https://noske.jerteh.rs i dostupni su ovlašćenim korisnicima. U ovom radu se za kolokacije određuje uzajamna leksičko-semantička „privlačnost“ na osnovu frekvencija i drugih mera u korpusima. Kolokacije se posmatraju u najširem smislu korpusne lingvistike - kao niz reči ili pojmova koji se pojavljuju zajedno, češće nego što bi se slučajno očekivalo. Predstavićemo kroz primere sedam glavnih tipova kolokacija: pridev + imenica (brza kontra), imenica + imenica (penal serija), glagol + imenica (postići gol), prilog + pridev (veoma talentovan), glagoli + predloška fraza (igra na stadionu) i glagol + prilog (šutirati snažno). Ekstrakcija kolokacija predstavlja tehniku računarske lingvistike za identifikaciju kolokacija u tekstu ili korpusu tekstova koristeći elemente slične rudarenju podataka, oslanjajući se na sintaksičke obrasce i frekvencije pojavljivanja.

Osim frekvencija pojavljivanja, razmatramo i druge faktore, poput bliskosti i konteksta u oba jezika. Na primer, da li određene kolokacije imaju specifična značenja ili se koriste samo u određenim situacijama. Takođe razmatramo da li su prethodno identifikovane kolokacije razumljive opštoj javnosti koja ne prati sport i nije upućena u jezik fudbala. Ukoliko ih prosečan govornik razume, govorimo o kolokacijama koje su postale deo javnog domena i nadmašile svoje poreklo fudbalskog domena.

Doprinos istraživanja čini i analiza veza između kolokacija i višečlanih termina. Veza je snažna kada višečlani termini sadrže kolokate sa jasnim značenjem unutar domena fudbala. Time pomažemo u razumevanju terminološke povezanosti unutar jezika fudbala, pružajući uvid u standardne kombinacije reči i njihovu upotrebu ilustrujući ih u fudbalskim korpusima srpskog i španskog jezika fudbala, što produbljuje njegovu analizu.

Ključne reči: *fudbal, korpusi, terminologija, kolokacije, srpski, španski*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Рада Стијовић, Ранка Станковић, Михаило Шкорић
Друштво за језичке ресурсе и технологије JePTex
E-mail: stijovicr@yahoo.com, {ranka|mihailo}@jerteh.rs

Речник савременог српског језика: РССЈ

Излагање ће представити мотивацију за израду Речника савременог српског језика (РССЈ), његов концепт и изазове у реализацији, као и примере неких решења, како из лексикографске перспективе, тако и из перспективе софтверског решења модела података и компоненти интегралног система. Друштво за језичке ресурсе и технологије JePTex прихватило је иницијативу удружења „Окупљени око језика" из дијаспоре и усмеравало га ка решењу које би, с једне стране, остварило жељене циљеве Удружења: израда папирне верзија речника и речничке базе података, а, с друге стране, било сагласно савременим лексикографским и информатичким стандардима.

Циљ подухвата је израда речника са око 50.000 речничких чланака, који ће представити лексику савременог српског стандардног језика. Грађа за речник је аутоматски ексцерпирана из електронских корпуса СрпКор2013 и СрпКор2021, као и из лексичке базе Лексимирка. У њој је заступљена лексика свих стилова стандардног језика (књижевноуметничког, научног, публицистичког, административног и разговорног) која се користи у последњих педесетак година, с тим што књижевноуметнички стил обухвата и нешто старија дела (од Другог светског рата наовамо). Лексика се обрађује методама модерне лексикографије, а објашњена значења дају се на начин приступачан и широкој читалачкој публици.

За развој РССЈ се користе различите методе засноване на подацима, које осим фамилије корпуса СрпКор и Лексимирке, укључују и, језичке моделе обучене у оквиру Друштва JePTex и пројекта ТЕСЛА (Text Embeddings - Serbian Language Applications). Модел података је инспирисан и у великој мери усаглашен са моделом Data Model for Lexicography (DMLex), док је база података имплементирана у систему PostgreSQL.

Изложићемо различите начине аутоматизације израде чланка које се ослањају на методе вештачке интелигенције и корпусне лингвистике, од фреквенција колокација, екстракције језичких образаца па све до аутоматског генерисања дефиниција и екстракције добрих примера употребе (GDEX, Good Dictionary Examples). Кориснички интерфејс лексикографске веб апликације развијене за потребе писања Речника, повезује речнички чланак са корпусима на платформи https://noske.jerteh.rs/. Прецизније речено, за одреднице речничког чланка или неку његову компоненту (израз, синоним, упућивање) лексикографска апликација прослеђује одговарајући CQL упит и параметри за тип обраде: конкорданце, колокације или фреквенције појављивања облика и израза. Коначно, представиће се развојно окружење речника и модалитети употребе.

Кључне речи: *речник, српски, лексикографија, лексикографска база, корпус*

Милена Милинковић
Институт за архитектуру и урбанизам Србије
E-mail: millena.milinkovic@gmail.com

Милица Иконић Нешић
Универзитет у Београду - Филолошки факултет
E-mail: milica.ikonic.nesic.fil@gmail.com

Именовани ентитети у дигиталном корпусу просторних планова

У излагању ће бити представљен поткорпус узорног доменског коруса просторног планирања и изазови на које се наилазило током његовог формирања и анотирања због атипичног садржаја. Овим поткорпусом је обухваћено шест планских докумената различитог просторног обухвата. Обрађени су текстови једног регионалног просторног плана, једног просторног плана јединице локалне самоуправе и четири просторна плана подручја посебне намене. Обрада и анотирање текстова, као и различити методолошки приступи спроведени су над поткорпусом коришћењем језичких модела, алата и ресурса за српски језик, развијених у оквиру Друштва за језичке ресурсе и технологије – JePTex.

Над припремљеним текстуалним садржајем планова, у формату чистог текста, извршено је сегментирање на реченице алатом Унитех, а потом, коришћењем постојећег Морфолошког електронског речника за српски језик, препознавање речи у тексту. Даља анотација поткорпуса просторних планова подразумевала је примену система за препознавање именованих ентитета SrpNER, односно аутоматско обележавање геополитичких појмова, назива организација и демонима. Прве две наведене класе именованих ентитета, даље су разврставане на потклассе, на пример, називи организација су разврстани на комерцијалне, политичке, верске и опште организације, односно организације које нису на други начин класификоване.

У даљем излагању ће бити представљен INCEpTION алат у оквиру ког је омогућено кориговање аутоматски анотираних именованих ентитета и њихово повезивање са базом знања Википодаци. Да би ово повезивање било могуће, било је потребно да за одговарајуће препознате именоване ентитете постоје ставке у википодацима. То је захтевало допуну базе знања креирањем ставки које су недостајале и употпуњавање већ постојећих ставки недостајућим подацима, односно својствима. Поред постављања различитих упита и излиставања именованих ентитета према унапред задатим критеријумима у оквиру система INCEpTION , креирањем SPARQL упита у бази знања Википодаци, омогућени су различити облици визуелизације добијених резултата, који између осталог, могу бити представљени у табеларном облику, у виду графа, као и географских карата.

Даља истраживања биће усмерена на допуну постојећег узорног корпуса просторног планирања и његовог поткорпуса просторних планова, као и на рад на побољшању и прилагођавању система за екстракцију именованих ентитета SrpNER, специфичностима планских докумената.

Кључне речи: *дигитални корпус, просторни планови, анотација корпуса, именовани ентитети, SrpNER, INCEpTION , Википодаци*

Захвалница: *Ово истраживање је подржао Фонд за науку Републике Србије, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Литература

[1] Bianchini, Carlo, Stefano Bargioni, and Camillo Carlo Pellizzari di San Girolamo. (2021). "Beyond VIAF: Wikidata As a Complementary Tool for Authority Control in Libraries". *Information Technology and Libraries* 40 (2). https://doi.org/10.6017/ital.v40i2.12959

[2] Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D., Stanković, R. (2021). Named Entity Recognition for Distant Reading in ELTeC. In Constanza Navarretta; Maria Eskevich (ed.), *CLARIN Annual Conference 2020,* 36-41. http://hdl.handle.net/10400.26/39114.

[3] Ikonić Nešić, M., Stanković, R., & Rujević, B. (2022). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca - Journal For Digital Humanities*, 21(2), 60-87. https://doi.org/10.18485/infotheca.2021.21.2.4

[4] Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation,* 24 (2): 473–489.

[5] Milinković, M. (2022). Application of TXM Tools for Spatial Plan Corpus Analysis. *Infotheca - Journal For Digital Humanities, 22*(1), 32-51. https://doi.org/10.18485/infotheca.2022.22.1.2

[6] Milinković, M. (2022a). *The development of library and language resources for organizing and finding information on spatial planning*. PhD dissertation. University of Belgrade - Faculty of Philology. https://hdl.handle.net/21.15107/rcub_raumplan_683

[7] Nielsen, F.Å., Mietchen, D., Willighagen, E. (2017). Scholia, Scientometrics and Wikidata. In: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds) *The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. Lecture Notes in Computer Science*. Vol 10577. Springer, Cham. https://doi.org/10.1007/978-3-319-70407-4_36

[8] Stanković, R. (2022). Distant Reading Training School 2020: Named Entity Recognition & Geo-Tagging for Literary Analysis. *Infotheca - Journal For Digital Humanities, 21*(2), 167_171. Retrieved from https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/223

[9] van Veen, Theo. (2019). Wikidata: From "an" Identifier to 'the' Identifier. *Information Technology and Libraries* 38 (2):72-81. https://doi.org/10.6017/ital.v38i2.10886.

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov
Univerzitet u Beogradu, Rudarsko-geološki fakultet
E-mail: {ranka.stankovic|jovana.radjenovic|maja.ristic|dragan.stankov}@rgf.bg.ac.rs

Kreiranje skupa za obučavanje modela za odgovaranje na pitanja na srpskom jeziku

Razvoj i primena veštačke inteligencije u jezičkim tehnologijama značajno su napredovali poslednjih godina, posebno u domenu zadatka odgovaranja na pitanja (Question Answering - QA). Dok su postojeći resursi za QA zadatke razvijeni za glavne svetske jezike, srpski jezik je relativno zanemaren u ovoj oblasti. Ovaj rad predstavlja inicijativu za kreiranje obimnog i raznovrsnog skupa podataka za obučavanje modela za odgovaranje na pitanja na srpskom jeziku, koji će doprineti unapređenju jezičkih tehnologija za srpski jezik.

Pored brojnih istraživanja o jezičkim modelima u poslednjih nekoliko godina, mnogo je urađeno i na referentnim skupovima podataka potrebnim za praćenje napretka modeliranja. Posebno je puno urađeno kada je reč o odgovaranju na pitanja i razumevanju pročitanog mada,

uglavnom, kada je reč o velikim jezicima (Rogers et al. 2023). U radu se pruža pregled različitih formata i domena raspoloživih višejezičnih i jednojezičnih resursa, sa posebnim osvrtom na srpski jezik (Cenić & Stojković 2023; Cvetanović & Tadić 2024). Razmotrićemo i implikacije koje slede iz prekomernog fokusiranja na engleski jezik

U okviru projekta TESLA (Text Embeddings - Serbian Language Applications) radi se na pripremi skupa podataka: kontekst, pitanja i odgovori, prikupljenih iz različitih domena. Skup će biti sačinjen od tri manja. U cilju izrade prvog skupa, podskup Stanfordovog skupa SQuAD (Rajpurkar et al. 2018), gde je odgovor segment teksta, prevodi se i prilagođava, odabirući teme kao što su: Nikola Tesla, klimatske promene, građevina, geologija, itd. Podskup će imati oko 7000 pitanja sa pratećim odgovorima. Drugi skup koji se priprema će uglavnom biti vezan za zaštitu životne sredine, informatiku i energetiku i sadržaće oko 5000 pitanja sa odgovorima i datim kontekstom ekscerpiranim iz udžbenika. Treći skup će sadržati automatski generisane kontekste na osnovu sadržaja baze znanja Wikidata.

Pitanja su pažljivo formulisana kako bi pokrila različite tipove upita: pitanja koja zahtevaju konkretne činjenice, pitanja sa deskriptivnim odgovorom (koja traže objašnjenja ili opis), i proceduralna pitanja, odnosno pitanja koja kao odgovor zahtevaju niz uputstava ili koraka. Podaci se prikupljaju na različite načine i verifikuju kroz proces ručnog anotiranja kako bi se obezbedila tačnost i relevantnost odgovora. Nedostatak ručno anotiranih skupova podataka na srpskom jeziku čini da doprinos ovog istraživanja bude od posebnog značaja.

Zaključak rada ukazuje na značaj i potencijal primene ovog skupa podataka u različitim oblastima, uključujući obrazovne tehnologije, digitalne asistente, i sisteme za pretragu informacija. Predstavljeni rezultati doprinose unapređenju jezičkih tehnologija za srpski jezik, i nadamo se da će podstaći dalja istraživanja i razvoj u ovoj oblasti.

Ključne reči: *veštačka inteligencija, obrada prirodnog jezika, jezički resursi, anotirani skupovi, ekstrakcija informacija, odgovaranje na pitanja*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Literatura

[1] Rogers, Anna, Matt Gardner, and Isabelle Augenstein. "QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension." ACM Computing Surveys 55, no. 10 (2023): 1-45. https://arxiv.org/pdf/2107.12708

[2] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." arXiv preprint arXiv:1806.03822 (2018). https://arxiv.org/abs/1806.03822 , https://rajpurkar.github.io/SQuAD-explorer/

[3] Cenić, Aleksandar B., and Suzana Stojković. "A Serbian Question Answering Dataset Created by Using the Web Scraping Technique." In 2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), pp. 147-150. IEEE, 2023.

[4] Cvetanović, Aleksa, Predrag Tadić, "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian", 2024, https://arxiv.org/html/2404.08617v1, https://paperswithcode.com/paper/synthetic-dataset-creation-and-fine-tuning-of

Јасмина Московљевић Поповић
Универзитет у Београду - Филолошки факултет
E-mail: jasmina.moskovljevic@fil.bg.ac.rs

Изазови при анотацији развојног корпуса

Рад се бави проблемима и изазовима на које се наилази приликом планирања и концептуализације стратегија и процедура које је потребно применити приликом анотирања развојног корпуса српског језика (РАКОПС-а). РАКОПС је репрезентативни развојни корпус писаног српског језика, креиран и сакупљен на Катедри за општу лингвистику и у Центру за примењену лингвистику Филолошког факултета Универзитета у Београду током последњих петнаестак година. Садржај РАКОПС-а, око 7000 текстова различитих жанрова чији су аутори ученици узраста 8 до 15 година, дигитализован је и доступан у два различита формата - .хтмл и .тхт, што га чини погодним за различите облике анализа.

Како анотирање развојног корпуса представља посебан изазов, било да је реч о ручном или аутоматском обележавању текста, у раду се представљају основни концепти који се односе на планиране процедуре и отвара се дискусија о различитим практичним питањима и потешкоћама на које се током овог процеса наилази.У сврху илустрације, наводе се конкретни примери из РАКОПСА, из текстова чији су аутори деца различитих узраста. У фокусу разматрања налази се анализа различитих процедура које је могуће применити на текстове који врло често у значајној мери одступају од утврђених ортографских конвенција, као и од уобичајених начина сегментације, а који уз то, врло често, врве од грешака. Разматрају се појаве уочене на свим централним нивоима лингвистичке анотације: (1) морфолошко тагирање (идентификација врста речи), (2) лематизација, (3) синтаксичко рашчлањавање, (4) семантичка анотација и (5) прагматска анотација.

Кључне речи: *писани језик, развојни корпус, РАКОПС; лингвистичка анотација, анотација грешака*

Саша Марјановић
Универзитет у Београду – Филолошки факултет
E-mail: sasa.marjanovic@fil.bg.ac.rs

Дејан Стошић
Универзитет Тулуз – Жан Жорес и CLLE (UMR 5263 CNRS)
E-mail: dejan.stosic@univ-tlse2.fr

Примена електронског конјугатора SerboVerb у проучавању овладавања глаголском флексијом српскога језика

Електронски конјугатор SerboVerb (Stosic i dr. 2018) представља један од језичких ресурса за српски језик настао на Универзитету Тулуз – Жан Жорес (Француска) у сарадњи с истраживачима с Филолошког факултета Универзитета у Београду. Ресурсу се бесплатно може приступити како путем мрежне апликације (на интернетској станици: https://serboverb.com), тако и помоћу мобилних апликација за телефоне типа Android и

Iphone. Намењен је првенствено корисницима који српским језиком овладавају као другим (било као страним, нематерњим, наследним, завичајним и др.), али се њиме могу служити и говорници свих узраста којима је српски језик матерњи. Ресурс се састоји од три модула (конјугацијски, речнички и игрификацијски), од којих су прва два интерно увезана, те чине језгро ресурса, а трећи представља збирку екстерно похрањених тестова за додатну примену прва два модула. Код првога — конјугацијског — модула реч је о претраживом исцрпном морфолошком лексикону у чијем је саставу 34.000 лематизованих глагола српскога језика с њима придруженим простим и сложеним флективним облицима, од којих је засад преко 20.000 ручно проверено и оверено. Код другога — речничког — модула глаголима из конјугацијског модула придодати су основни еквиваленти на другим језицима. При томе, за фонд од 1.800 глагола којима се овладава на вишим нивоима према Заједничком европском оквиру за језике (ЗЕРОЈ) тренутно су укључени еквиваленти на десет језика, а за стотину најчешће претраживаних глагола у конјугацијском модулу унети су еквиваленти на укупно 36 језика.

Функције су целокупног ресурса, стога, вишеструке (уп. Marjanović i dr. 2023): прво, од користи може бити у комуникацијским ситуацијама, с једне стране при разумевању свакога појединачног облика глагола и његовог повезивања с канонским обликом, с друге при производњи било којег облика у парадигми глагола, полазећи од било којег кориснику познатог облика, али се такође може користити и у сазнајним ситуацијама при учењу и усвајању појединачних флективних облика и глаголске флексије српскога језика уопште. С обзиром на то што мрежна апликација овог ресурса садржи административни панел помоћу којег се прикупљају и прегледају разноврсни подаци о употреби ресурса, администраторима је на располагању могућност да прикупљене податке примене на извођење закључака који могу имати значаја при осмишљавању дидактизовања глаголске флексије српскога језика као другог. У овоме излагању понајпре се показује како учесталост претраживања одређених глагола, укрштена с подацима о њиховим фреквенцијама из општег корпуса српског језика и глаголским пописима из уџбеника српскога језика као страног, оцртава тенденције расподеле глагола према нивоима ЗЕРОЈ-а; затим, показује се на који начин учесталост претраживања глагола и њихових флективних облика истиче примарне конјугацијске обрасце за овладавање у српскоме као другом; коначно, указује се, индиректним путем, на потешкоће које корисници евентуално имају у овладавању глаголском флексијом српског језика.

Кључне речи: *SerboVerb, глагол, флексија, конјугатор, овладавање другим језиком*

Технологије обраде јужнословенских језика

Marija Đokić Petrović
Virtual Vehicle Research GmbH
Inffeldgasse 21a, 8010, Graz, Austria
E-mail: office@marijadjokicpetrovic.com

Mihailo St. Popović
Austrian Academy of Sciences, Institute for Medieval Research
Georg-Coch-Platz 2, 1010, Wien, Austria
E-mail: mihailo.popovic@oeaw.ac.at

Vladimir Polomac
University of Kragujevac, Faculty of Philology and Arts
Jovana Cvijića bb, 34000, Kragujevac, Serbia
E-mail: v.polomac@filum.kg.ac.rs

Korišćenje prepoznavanja imenovanih entiteta za analizu srpskih arhivskih dokumenta

Digitalne humanističke nauke predstavljaju značajan napredak u istraživanju i očuvanju kulturnog nasleđa, omogućavajući istraživačima da koriste savremene tehnologije za analizu i tumačenje istorijskih dokumenata. U našem radu analiziramo dva srpska arhivska dokumenta pisana brzopisanom ćirilicom. Prvi dokument iz 1778. godine je neobjavljen i čuva se u arhivi grčke crkve Svetog Đorđa u Beču, najstarije pravoslavne crkve u današnjoj Austriji. Drugi dokument iz 1500. godine je testament Miloša Belmuževića, srpskog vlastelina u službi ugarskih kraljeva. Dokumenti su u početku bili podvrgnuti ručnom čitanju i ispitivanju kako bi se izvukle informacije o osobama, lokacijama i demografiji. Istovremeno, za postizanje istog zadatka korišćen je računarski pristup koji koristi Prepoznavanje Imenovanih Entiteta (NER). Za obuku NER modela korišćen je skup podataka koji je sadržao Povelju kralja Stefana Uroša II Milutina manastiru Svetog Stefana u Banjskoj s početka 14. veka, Dečansku hrisovolju iz 14. veka i zbirku povelja i pisama iz 13. veka iz Dubrovačkog arhiva. Na kraju je izvršena komparativna analiza tradicionalnog pristupa i rezultata računarske obrade, olakšavajući procenu efikasnosti obe metode. Ovom analizom je dodatno potvrđena vrednost Digitalnih humanističkih nauka u očuvanju i proučavanju srpskih istorijskih dokumenata.

Ključne reči: *Prepoznavanje Imenovanih Entiteta; Digitalne humanističke nauke; Brzopisana ćirilica; Kulturno nasleđe; Srpski arhivski dokumenti; Istorijsko nasleđe*

Literatura

[1] Todorović, B. Š., Krstev, C., Stanković, R., & Nešić, M. I. (2021, September). *Serbian NER&Beyond: The archaic and the modern intertwinned*. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 1252-1260).

[2] Cvejić, A. (2022). *Analiza različitih modela za Prepoznavanje Imenovanih Entiteta na srpskom jeziku*. Zbornik radova Fakulteta tehničkih nauka u Novom Sadu, 37(02), 316-319.

[3] Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). *A survey on Named Entity Recognition—datasets, tools, and methodologies*. Natural Language Processing Journal, 3, 100017.

Nikola Janković
Filološki fakultet, Univerzitet u Beogradu
E-mail: nikolajankovickv@gmail.com

Metodologija izrade višejezičnog paralelnog korpusa na osnovu onlajn digitalnih uputstava za upotrebu: korpus Hilti uputstava

U radu je predstavljena metodologija izrade opsežnog višejezičnog paralelnog korpusa uputstava za upotrebu na osnovu digitalnih uputstava za upotrebu uređaja kompanije Hilti. Cilj ovog rada je doprinos boljim resursima za istraživanja vezanim za mašinsko prevođenje, teoriju prevođenja, ali i za nastavu prevođenja i druge vrste lingvističkih istraživanja. Polazna tačka metodologije bilo je korišćenje uniformne strukture HTML uputstava za svaki jezik na zvaničnoj Hilti onlajn bazi uputstava. Dalji koraci su obuhvatali identifikovanje HTML elemenata koji sadrže relevantne delove teksta, njihovu doslednu numeraciju za svaki jezik, identifikovanje i čišćenje pogrešno paralelizovanih tekstova i čuvanje obrađenih tekstova u TMX formatu, tako da je engleski izvorni jezik u svakom fajlu. U proseku, za svako uputstvo postoji oko 20 ciljnih jezika, a sam korpus sadrži oko 7 miliona reči. Prednosti kreiranog korpusa uključuju visoku preciznost kada je u pitanju podudaranje segmenata između jezika, značajan broj reči, kao i veliki broj jezika obuhvaćenih korpusom. U radu će biti detaljno opisani koraci preuzimanja, obrade i strukturiranja podataka, tehnički izazovi sa kojima smo se suočili i kako su rešeni, kao i osnovni statistički podaci vezani za jezike u korpusu. Smatramo da se ovakav tip korpusa može koristiti za unapređivanje resursa za prevođenje, teoriju prevođenja, nastavu prevođenja, kao i za različite vrste analiza stručnog jezika.

Ključne reči: *paralelni korpusi, mašinsko prevođenje, jezički alati, stručni jezik, TMX*

Анђелка Зечевић, Анастазиа Жунић, Кристина Милојевић
Математички институт Српске академије наука и уметности
E-mail: {andjelkaz|anastazia.zunic|kristinam}@mi.sanu.ac.rs

Стари текстови, нове технологије: дигитализација докумената на српском језику

У свету дигиталне хуманистике појављују се многобројни алати и платформе које имају за циљ да унапреде и обогате рад са дигитализованом грађом. У основи ових алата су комплексни алгоритми машинског учења за обраду слика и текста обучавани над подесно припремљеним репозиторијумима докумената, док су њихове најчешће функционалности унапређивање квалитета слика, просторна анализа докумената, оптичко препознавање карактера и корекција рашчитаног садржаја. Новији алати нуде подршку и за стандардизовање ортографије, лакше и свеобухватније претраживање докумената, издвајање топонима, препознавање тема у садржајима и генерисање сажетака докумената.

У овом раду биће представљена искуства у коришћењу модерних алата у отвореном приступу за дигитализацију докумената на српском језику. У питању су алати

Calamari-OCR, docTr, LayoutParser, Kraken, Tesseract, OCR4All и други, дизајнирани за појединачне кораке процеса дигитализације или потпун процес пратећи парадигму са-краја-на-крај. За тестирање алата ће се користити грађа Народне библиотеке Србије и периодике објављиване у току деветнаестог века које карактерише велика разноврсност графичких елемената, нестандардни формати, али и физичка деградација и лошији квалитет скенираних докумената. Уз уочене предности и ограничења ових алата, биће дискутовани и начини којима се ови алати даље могу проширивати и прилагођавати српском језику. Посебан осврт ће бити на улози језичких технологија и сценаријима у којима је њихово коришћење више него потребно.

Кључне речи: *дигитализација докумената, периодике, обрада слика, обрада природног језика*

Nikitas N. Karanikolas, Professor
Department.of Informatics & Computer Engineering, University of West Attica, Greece
E-mail: nnk@uniwa.gr

Екстракција именичких и предложних фраза за грчки језик библиотеком Spacy

Разумевање природног језика (Natural Language Understanding, NLU) обухвата различите задатке. Кратак сажетак овог процеса подразумевао би да реченице из текста: а) буду синтаксички анализиране, б) да фразама које конституишу синтаксичку структуру (*phrase structure*) буду приписане одговарајуће семантичке/тематске улоге које кореспондирају са доступним позицијама у глаголском оквиру (*frame*, у терминима Филморове семантике оквира). Фразе које се налазе на овим позицијама и којима је потребно приписати одговарајуће семантичке/тематске улоге су именичка фраза (NP) и предлошка фраза (PP). Сходно томе, парсер (синтаксички анализатор, анализатор фразне структуре), било да је реч о плитком или дубоком рашчлањивању (*shallow/deep parsing*), треба да буде у стању да екстрахује NP и PP фразе.

Алтернативни приступ овом проблему подразумевао би да се у првом кораку (синтаксичкој анализи) користи анализатор заснован на депенденцијалној теорији, те да се за сваку реченицу креирају депенденцијална стабла, односно стабла зависности. Депенденцијална стабла кодирају односе зависности између свих речи и фраза које сачињавају реченицу. На врху стабла (у његовом *корену*, у терминима депенденцијалне граматике) налази се реченични предикат, од кога зависе сви остали реченични чланови. Депенденцијално стабло приказује управне речи изнад речи које су од њих зависне, док се изнад лукова који их повезују означавају граматичке функције. Предност депенденцијалних дијаграма у односу на дијаграме фразне структуре је у томе што су предикат и његови аргументи директно повезани.

Ресурси неопходни за обраду природних језика су велики и захтевни. Да поменемо само неке: лексикон који препознаје све облике флективне речи и може да екстрахује њену лему, као и релевантна морфосинтаксичка обележја (род, број и падеж код именица; време, начин, стање, лице и број код глагола, итд.), затим попис глаголских оквира (са специфікованим тематским улогама и пописом селекционих рестрикција којима се ограничава избор аргумената, итд.).

Иако постоје алати који омогућавају анализу фразне структуре или депенденцијалну анализу и без употребе одговарајућег лексикона, није јасно да ли су ови алати у стању да изврше адекватну анализу фразне структуре, као и да издвоје релевантне NP и PP које су носиоци одговарајућих тематских улога и подлежу селекционим рестрикцијама које дефинише дати глаголски оквир. У овом раду се евалуира један такав алат за грчки језик – библиотека SpaCy.

Кључне речи: *грчки језик, екстракција NP и PP, депенденцијално рашчлањивање, SpaCy*

Литература

[1] Daniel Jurafsky & James H. Martin, Speech and Language Processing, 2019. Chapter 15 Dependency Parsing. https://web.stanford.edu/~jurafsky/slp3/old_oct19/15.pdf
[2] Fei Xia & Martha Palmer, Converting Dependency Structures to Phrase Structures. Human Language Technology - The Baltic Perspectiv, 2001. https://aclanthology.org/H01-1014.pdf
[3] Nikitas Karanikolas et al, Large language models versus natural language understanding and generation. ACM Digital Library, pub. 14 Feb. 2024. https://doi.org/10.1145/3635059.3635104

Jaka Čibej
Faculty of Arts, University of Ljubljana
Centre for Language Resources and Technologies, University of Ljubljana
E-mail: jaka.cibej@ff.uni-lj.si

Prvi koraci ka onlajn servisu za automatsku morfološku fleksiju srpskog i hrvatskog

Mašinski čitljivi morfološki leksikoni otvorenog koda korisni su za morfosintaksičko označavanje korpusa i predstavljaju ključni korak ka sastavljanju savremenih baza podataka digitalnih rečnika (npr. Kosem et al. 2021). Među leksikonoma za južnoslovenske jezike trenutno je najrazvijeniji *Morfološki leksikon slovenačkog jezika Sloleks* (Čibej et al. 2022). Verzija 2.0 sa približno 100.000 unosa je ažurirana na verziju 3.0, dodajući približno 265.000 novih unosa, njihovih oblika sa naglaskom, naglašenih oblika i IPA/SAMPA izgovora. Svi su automatski generisani koristeći *Pregibalnik*, prilagođeni alat otvorenog koda (takođe dostupan kao API servis[1]) za proširenje slovenačkog leksikona, koji uzima lemu i njene morfosintaksičke karakteristike u skladu sa morfološkim specifikacijama MULTEXT-East[2] (npr. *liofilizacija*, imenica, zajednička, ženskog roda) kao input i generiše (između ostalog) potpune paradigme oblika – padež, broj, vreme itd. (npr. liofilizacija, liofilizacija, ...) kombinacija mašinskog učenja i lingvistički informisanih metoda zasnovanih na pravilima, uključujući mašinski čitljive morfološke obrasce (npr. „[*liofilizacij*]-a, [*liofilizacij*]-e, [*liofilizacij*]-i, ...“) koji su automatski izvučeni i potvrđeni (Arhar Holdt & Čibej 2018; Arhar Holdt 2021) pre nego što se koriste u predviđanjima mašinskog učenja.

Objavljena su dva leksikona otvorenog koda slična Sloleksu za srpski i hrvatski – srLex 1.3 (Ljubešić 2019a) i hrLex 1.3 (Ljubešić 2019b), sastavljena iz srWaC i hrWaC korpusa. Međutim, dok su metode mašinskog učenja za proširenje leksikona već bile korišćene za predviđanje paradigmi za hrvatski i srpski jezik (npr. Ljubešić et al. 2016; takođe Šnajder 2013),

paradigme su bile dostupne samo u formatu Apertium,[3] koji nije kompatibilan sa infrastrukturom *Pregibalnika*. Obrasce je potrebno konvertovati i unakrsno proveriti sa srLexom i hrLexom da bi bili uspešno implementirani u lako dostupnom API servisu. Pošto su hrvatski i srpski strukturno slični slovenačkom i dele sličan infrastrukturni okvir, isti metod primenjen na slovenačke podatke može se koristiti (uz neka manja prilagođavanja) za implementaciju obrazaca u *Pregibalnik*. Ovo neće biti samo prvi korak ka proširenju funkcionalnosti *Pregibalnika* tako da pokrije srpski i hrvatski jezik i pomoći će automatskom proširenju leksikona novim unosima, već će takođe identifikovati potencijalne nedoslednosti u trenutnim verzijama leksikona.

Priznanje: *Autor se zahvaljuje na finansijskoj podršci Slovenačke agencije za istraživanje i inovacije (osnovno finansiranje istraživanja br. P6-0411 – Jezički resursi i tehnologije za slovenački jezik).*

Ključne reči: *leksika, morfologija, fleksija, proširenje, hrvatski, srpski*

Literatura

[1] Arhar Holdt, Špela & Jaka Čibej, 2018, Oblikoslovni vzorci v leksikonu Sloleks: izhodiščni nabor za samostalnike. In: *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave, 6(2)*, pp. 33-66. https://doi.org/10.4312/slo2.0.2018.2.33-66

[2] Arhar Holdt, Špela, 2021. Oblikoslovni vzorci za strojno procesiranje slovenščine. In: Arhar Holdt, Špela (ed.): *Nova slovnica sodobne standardne slovenščine: viri in metode*. Založba Univerze v Ljubljani. https://doi.org/10.4312/9789610605478

[3] Čibej, Jaka, Kaja Gantar, Kaja Dobrovoljc, Simon Krek, Peter Holozan, Tomaž Erjavec, Miro Romih, Špela Arhar Holdt, Luka Krsnik, Marko Robnik-Šikonja, 2022, *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1745

[4] Kosem, Iztok, Simon Krek, Polona Gantar, 2021, Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian In: *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion*, pp. 81–83.

[5] Ljubešić, Nikola, 2019a, *Inflectional lexicon srLex 1.3*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1233.

[6] Ljubešić, Nikola, 2019b, *Inflectional lexicon hrLex 1.3*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1232.

[7] Ljubešić, Nikola, Filip Klubička, Željko Agić, 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'16)*, pp. 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

[8] Šnajder, Jan, 2013, Models for predicting the inflectional paradigm of Croatian words. In: *Slovenščina 2.0, 1 (2)*, pp. 1–34. http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_02.pdf.

Martina Pavić
Institut za hrvatski jezik
E-mail: mpavic@ihjj.hr

Frekvencija značenja pridjeva u hrvatskome medicinskom nazivlju – korpusno utemeljena analiza kolokacija

Pridjevi su važna sastavnica hrvatskoga medicinskog nazivlja koja pridonosi njegovoj visokoj strukturiranosti sužavajući i modificirajući značenje riječi kojima se pridijevaju te mogu biti presudni u formiranju pojmovnoga sustava (Pitkänen-Heikkilä 2015, vidi Pavić 2022: 153).

Frekvencija značenja pridjeva u medicini odražava način na koji se opisuju zdravstveno stanje, dijagnoze i tretmani. Istraživanje frekvencije ključno je za standardizaciju i preciznost medicinskoga nazivlja, što je iznimno važno za jasnoću i sigurnost u medicinskoj komunikaciji i praksi. U medicinskome području ta vrsta riječi sudjeluje u klasifikaciji i specifikaciji bolesti, dijagnoza, simptoma, indikacija i drugih pojmovnih kategorija. Smatrajući da značenjska ostvarenja pridjeva u medicinskome nazivlju treba promatrati u uporabi, u radu ćemo primijeniti korpusni model istraživanja, pri čemu ćemo se služiti znanstvenim korpusom suvremenih medicinskih časopisa, veličine 5 318 395 pojavnica, koji je sastavljen u alatu *Sketch Engine* (Pavić 2022).

U radu ćemo primijeniti klasifikaciju na 14 značenja pridjeva u hrvatskome medicinskom nazivlju – *funkcija*, *uzrok*, *vrijeme/trajanje*, *fizička značajka*, *sastav*, *stanje*, *svojstvo*, *lokacija*, *ishod*, *primjena*, *izvor*, *dob*, *intenzitet*, *odnos/relacija* – predloženu u Pavić 2022. U značenjskoj analizi poći ćemo od vršnih kategorija naziva, odnosno primijenit ćemo model odozgo prema dolje (engl. *top-down approach*). Od Taylora (2017) preuzete su kategorije naziva <bolest>, <simptom>, <indikacija>, <dijagnoza> i <postupak>, kojima su, nakon iscrpna pretraživanja korpusa, dodani <medicinsko osoblje>, <bolesnik>, <nuspojava>, <dio tijela> i <medicinska oprema>. Taylorovu kategoriju naziva <lijek> uključit ćemo u kategoriju naziva <terapija> jer se *terapija* najčešće leksikalizira imenicom *lijek* te vrstama lijeka (Pavić 2022).

Prvi ćemo put istražiti frekvenciju značenja pridjeva u hrvatskome medicinskom nazivlju koji su izlučeni poluautomatskom metodom u opciji *Keywords* (broj pojavnica veći od 200). Pritom ćemo poći od popisa najčešćih pridjevsko-imeničkih kolokacija (Pavić 2022) jer je pretpostavka da pridjev izriče različite značajke ovisno o imenici koju modificira (Grčić Simeunović 2021). U kolokacijama ćemo utvrditi koje značenje odnosno značenja pridjevi izražavaju (primjerice pridjev *hormonski* izražava četiri značenja u različitim pridjevsko-imeničkim svezama). Osim pridjeva inherentno medicinskoga značenja (*krvni*, *endoskopski*, *gastrointestinalni*…) istražit ćemo i frekvenciju značenja pridjeva iz općega jezika koji u višerječnim svezama tvore medicinske nazive (*desni*, *dubok*, *središnji*…).

Cilj je rada primjenom korpusne metode istraživanja utvrditi najčešća značenja pridjeva u hrvatskome medicinskom nazivlju, propitati koji pridjevi u kolokacijama mijenjaju kategoriju naziva i mijenja li se pritom i njihovo značenje. Uključivanjem pridjeva iz općega jezika u značenjsku analizu ujedno će se propitati odnos općega i specijaliziranoga jezika te postojanje eventualnoga pomaka u značenju.

Ključne reči: *pridjevi, značenjska analiza, hrvatsko medicinsko nazivlje, znanstveni korpus, korpusna metoda, kolokacije*

Literatura

[1] Grčić, Simeunović, L. (2021) *Terminološki opis u službi stručnoga prevođenja. Dinamično modeliranje specijaliziranoga znanja*. Zadar – Zagreb: Sveučilište u Zadru – Institut za hrvatski jezik i jezikoslovlje.

[2] Pavić, M. (2022) *Uloga pridjeva u hrvatskome medicinskom nazivlju*. Doktorska disertacija. Sveučilište u Zagrebu, Filozofski fakultet, Zagreb.

[3] Pitkänen-Heikkilä, K. (2015) Adjective as terms. *Terminology*, 21 (1), 76–101.

[4] Taylor, B. R. (2017) *The Amazing Language of Medicine: Understanding Medical Terms and Their Backstories*. New York City: Springer International Publishing.

Ana Ostroški Anić, Ivana Brač
Institute for the Croatian Language, Zagreb, Croatia
E-mail: aostrosk@ihjj.hr

Opis glagola mišljenja u hrvatskom prema semantici okvira

Semantički opis glagola u različitim jezičnim resursima uglavnom uključuje podjelu glagola u semantičke skupine, obično slijedeći pionirski rad Levin (1993), bez temeljitije analize.

Istraživački projekt SEMTACTIC (https://semtactic.jezik.hr/) nastoji nadmašiti svoj temeljni cilj određivanja semantičkih skupina 500 najčešćih glagola u hrvatskom jeziku istraživanjem njihovih prototipnih sintaktičkih obrazaca i semantičkih uloga. Odnos između semantike i sintakse hrvatskih glagola istraživat će se kako unutar semantičke skupine, tako i između skupina različitim opisima, od kojih je jedan teorijski pristup korišten za definiranje glagola prema načelima semantike okvira (Fillmore 1985; Fillmore, Johnson i Petruck, 2003; Ruppenhofer i dr., 2016).

Ovaj rad predstavlja opis glagola mišljenja u bazi podataka Verbion prema teoriji semantike okvira, koji se razvija u okviru projekta. Objašnjena je argumentna struktura glagola poput misliti, smisliti, promisliti, razmišljati, predložiti, itd., zajedno s njihovim semantičkim opisom unutar odgovarajućih semantičkih okvira FrameNeta, npr. Mišljenje, Procjenjivanje ili Smišljanje. Relevantni korpusni primjeri rečenica u kojima se ti glagoli nalaze kao ciljne leksičke jedinice označeni su prema elementima okvira kako bi se usporedili glagoli prema njihovim valencijskim okvirima. Npr., u rečenici [Ovu inovativnu uslugu]Idea SMISLILA je [tvrtka iz Arizone koja proizvodi prirodne preparate za njegu tijela]Cognizer označeni su nužni elementi okvira Smišljanje.

Definiranje glagola mišljenja prema metodologiji korištenoj u FrameNetu (Ruppenhofer i dr., 2016) daje dodatnu razinu sintaktičkoga i semantičkoga opisa u bazi podataka Verbion, kao i omogućuje podatke za stvaranje budućega leksikona hrvatskoga jezika temeljenoga na okvirima.

Ključne reči: *glagoli, glagoli mišljenja, FrameNet, semantika okvira*

Literatura

[1] Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di semantica: Rivista internazionale di semantica teorica e applicata* 6, 222–254.

[2] Fillmore, Charles J.; Johnson, Christopher R.; Petruck, Miriam R. L. 2003. Background to Framenet. *International journal of lexicography* 16/3, 235–250.

[3] Levin, Beth. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. The University of Chicago Press.

[4] Ruppenhofer, J. et al. 2016. *FrameNet II: Extended Theory and Practice*. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf

Marija Pantić
JeRTeh
E-mail: p.mmmmmm@gmail.com

Veštački korpusi za mašinsku obuku alata za proveru gramatike

Mašinska obuka alata za proveru gramatike zahteva velike korpuse gramatički ispravnih i neispravnih rečenica. Prikupljanje prirodno nastalih negramatičnih rečenica i njihova klasifikacija prema broju i vrsti grešaka mogu biti komplikovan, dugotrajan, i zahtevan posao s obzirom na potrebne ljudske i jezičke resurse. Veštački korpusi grešaka mogu biti relativno jednostavna i znatno jeftinija, a često i daleko pouzdanija, alternativa, a mogu se proizvesti indukcijom grešaka minimalnim transformacijama u korpusima gramatički ispravnih rečenica sa odgovarajućim morfosintaksičkim oznakama. Korpusi gramatičnih rečenica mogu biti opšti ili namenski sastavljeni za potrebe indukcije grešaka. Opšti korpusi ne zahtevaju dodatnu pripremu, ali moraju sadržati dovoljno primera relevantnih obrazaca, a dobijeni korpus indukovanih negramatičnih rečenica biće manji od izvornog korpusa gramatičnih rečenica. Nasuprot ovome, odabir rečenica koje sadrže samo obrasce relevantne za indukciju odabranog tipa greške omogućava transformaciju svake gramatične rečenice iz izvornog korpusa u negramatičnu rečenicu u indukovanom korpusu u skladu sa pravilima indukcije. U odabiru tipa grešaka za indukciju i pisanju indukcionih pravila mora se imati u vidu sledeće:

1. Indukcija greške ne mora uvek proizvesti negramatičnu rečenicu.

   a) *Pričam.* → *Priča.*
   b) *Pričam.* → *Pričah.*
   c) *Ja pričam.* → **Ja priča.*
   d) *Pričam koliko hoću.* → *?Priča koliko hoću.*

U primerima (1a) i (1b) zamena prvog lica jednine bilo kojim drugim licem ili prezenta drugim prostim ličnim glagolskim oblikom neće proizvesti negramatičnu rečenicu jer se rečenica sastoji samo iz glagola. U primeru (1c) uspešno smo indukovali neslaganje subjekta i predikata u licu. U primeru (1d) proizveli smo gramatički moguću ali sa stanovišta upotrebe upitnu rečenicu.

2. Negramatičnu rečenicu često je moguće ispraviti na nekoliko načina primenom istog minimalnog broja transformacija.

   a) *On je pevao.* → *On pevao.* → *On je pevao; On bi pevao; On beše pevao: On peva; On pevaše.*
   b) *On je pevao.* → *On je pevala.* → *Ona je pevala; On je pevao.*

U primerima (2a) i (2b) ispravna rečenica ne nudi jedino a ni očigledno najbolje rešenje za ispravku greške u neispravnoj rečenici.

U ovom izlaganju biće razmotreni neki od kriterijuma za ocenu kvaliteta indukovanog korpusa, preterana i nedovoljna specifičnost indukcionih pravila za različite tipove gramatičkih grešaka u srpskom, i neke od najučestalijih grešaka izvornih i stranih govornika srpskog koje bi mogle biti bolje obuhvaćene alatima za proveru gramatike.

Ključne reči: *gramatički alati, veštači korpusi, mašinsko učenje*

Граматика и лексика јужнословенских језика у контексту обраде природних језика

Марина Баги
Институт за српски језик САНУ
E-mail: marina.bagi@isjsanu.ac.rs

Синтаксичка и семантичка анализа глагола *оштетити* и *уништити* из угла теорије семантике оквира

У овом раду представља се анализа глагола *оштетити* и *уништити* из угла теорије семантике оквира. Полазећи од претпоставке да синтаксички феномени корелирају са семантичким (и обрнуто), за теоријску основу рада бира се теорија семантике оквира, когнитивносемантички приступ Ч. Филмора који експлицитно повезује значења неке речи са синтаксичким контекстима у којима се она јавља. Основна теза ове теорије је да значења неке сублексеме треба посматрати у вези са семантичким оквирима, који представљају схематске приказе појмовних структура и образаца веровања који леже у основи значења речи. Ова теорија у основи је Фрејмнета, електронске лексикографске базе података која је заснована на корпусу, а у оквиру које су приказани оквири које различите лексеме призивају. Наиме, једно од централних начела овог приступа јесте да одређене врсте речи (глаголи, именице, придеви, прилози) призивају оквире, својеврсне системе појмова повезаних на такав начин да је за разумевање једног појма потребно разумевање целе структуре чији је он део.

У овом истраживању посматра се више од пет стотина примера употребе глаголских лексема *оштетити* и *уништити* ексцерпираних из електронског *Корпуса савременог српског језика* (на платформи https://noske.jerteh.rs/) са циљем да се опише и испита њихово синтаксичко и семантичко понашање, те да се прикаже њихова употреба. Стога се детаљно описују оквири које призивају поменути глаголи, као и елементи оквира, а изводе се и закључци у вези са синтаксичко-семантичким интерфејсом. Уз то, будући да су у питању глаголи сличне семантике, циљ је да се истраже сличности и разлике између њих, што се чини употребом алата доступних на платформама *Sketch Engine* и *No Sketch Engine*. Проучавају се њихови најчешћи колокати, фреквенције неких облика, скице речи, синоними и сл. Спроведено истраживање део је будућег, ширег истраживања у оквиру рада на докторској дисертацији која се бави применом теорије семантике оквира у српској лексикографији (на примеру глагола оштећивања и уништавања), као и део пројекта израде српског Фрејмнета. Један од циљева истраживања је и да се опишу изазови који су се јављали током рада на овом делу пројекта.

Кључне речи: *теорија семантике оквира, синтакса, семантика, глагол оштетити, глагол уништити, обрада природних језика, српски језик*

Проф. др Наташа Киш
Универзитет у Новом Саду, Филозофски факултет, Одсек за српски језик и лингвистику
E-mail: natasab@ff.uns.ac.rs

Синтаксичко-семантичка анотација електронских корпуса српског језика

У раду ће бити указано на могућности коришћења електронског корпуса као примарног извора грађе за различита истраживања из области синтаксе и семантике савременог српског језика. Два основна задатка рада јесу да се прикажу досадашња истраживања везана за процес комплементације придева и придевских именица, односно њиховог рекцијског потенцијала на синтагматском плану, а која се почивају на релевантој грађи преузетој првенствено из електронског *Корпуса савременог српског језика*, СрпКор2013 (www.korpus.matf.bg.ac.rs). Само приликом анализе процеса допуњавања придева из корпуса је ексцерпирано преко 9000 реченица. Други задатак је да се из угла анализе на нивоу просте реченице укаже на потребу проширења могућности претраге корпуса. Претраге овог типа укључивале би податке о валенцијским особинама предиката реченице и свим обавезним и факултативним актантима у конкретној ситуацији, а важан сегмент анализе јесте издвајање семантичких класа речи и њихових синтаксичко-семантичких особености на основу којих би се могли претраживати различити реченични модели.

На пример, релациони придев *веран* везује уз себе објекатску допуну у дативу без предлога са ознаком живо+/- (он је *веран* ***пријатељу*** / ***тој идеологији***), именица *верност* добија исти тип објекатске допуне реализоване различитим формама (показао је *верност* ***пријатељу*** / ***према пријатељу***), али се може у одређеном контексту употребити и без допуне (они не могу да говоре о *верности*).

Посебну пажњу треба скренути на вишезначност лексема која има утицај и на рекцијски потенцијал речи. Придев *веран* и именица *верност*, уколико носилац особине има ознаку живо-, могу имати значење „бити једнак нечему“, што имплицира да ће допуна коју регирају припадати другачијем семантичком типу, односно имаће значење другог носиоца особине (текст је *веран* ***оригиналу***; издавач се држао *верности* ***према тексту***).

Претрага корпуса била би прецизнија уколико би се као параметри могли укључити различити синтаксичко-семантички подаци као што су аниматност, усмереност / директивност, реципрочност, аблативност, обухватност, партитивност и сл. У структурирању реченице, релевантно је и питање којим се језичким средствима исказују подаци о агенсу/псеудоагенсу. У савременом српском језику, на пример, издвајају се различити структурни типови с обзиром на ове параметре (***ја*** *бих јела јабуке* / *једу* ***ми*** *се јабуке* и сл.).

Из угла говорника српског као матерњег или као нематерњег/страног језика било би корисно да се у претрагама електронског корпуса може поћи од семантичких класа речи чију би формализацију, односно структурне моделе у којима се реализују показали примери ексцерпирани из електронског корпуса (нпр. у класи речи са значењем оптативне модалности налазе се придев *жељан*, именица *жеља* и глагол *желети*, који се могу реализовати у следећим реченичним моделима: Ја сам *жељна* ***одмора*** – Моја *жеља* ***за одмором*** је велика / Ја имам *жељу* ***за одмором*** – Ја *желим* ***да се одморим / одмор***).

Наведени параметри претраге припадају домену семантичко-синтаксичке анотације електронских корпуса, те је од велике важности и указати на одговарајуће лингвистичке анализе које би омогућиле издвајање и укључивање потребних података у саме корпусе.

Кључне речи: *синтакса, семантика, анотација еликтронског корпуса, рекција, српски језик*

Таня Нейчева
Пловдивски университет „Паисий Хилендарски“
E-mail: neyche1va@uni-plovdiv.bg

Ворителен предикативен падеж в съвременния сръбски, руски и полски език (по данни от многоезичен онлайн речник)

Съществуващите днес паралелни езикови корпуси дават възможност за съпоставителни изследвания върху материал от два (рядко повече) езика. За съжаление, многоезичните корпуси все още не са особено богати и разчитат най-вече на отделни преводи на художествена литература. В опит да намери друг надежден източник на материал за съпоставителни изследвания авторката се обръща към многоезичните онлайн речници. Този тип платформи разчитат на т. нар. translation memory – постоянно допълвана база данни от преводни текстове, която позволява търсената в речника дума (или цяло словосъчетание) да се провери в многобройни контексти на оригиналния и на преводния език и дори паралелно в няколко езика. В този смисъл основната характеристика, която ги отличава от типичните езикови корпуси, е липсата на граматична анотация (каквато впрочем не притежават и част от корпусите).

Представеното изследване е посветено на славянския творителен предикативен падеж и е осъществено върху ексцерпирани от многоезичния онлайн речник Glosbe паралелни примери от сръбски, руски и полски език, като в хода на работата беше изпробван прост алгоритъм за ръчно откриване на необходимата граматична информация в неанотиран корпус. Основните стъпки включват: подбор на най-адекватните ключови думи за търсенето, извличане на примерите, отстраняване на нерелевантните и анотиране на релевантните резултати, анализ на данните.

В резултат от изследването бяха описани особеностите в употребата на творителния предикативен падеж и конкуренцията му със съгласувателните предикативни (именителен, дателен и отчасти винителен) падежи в съвременния сръбски, руски и полски език при няколко основни подтипа на съставното именно сказуемо – с лична, нелична и безлична форма на копулата; с преходна и непреходна (възвратна) лична полукопула и с нелична форма на полукопулата. Бяха откроени както разликите, така и някои неописани досега сходства между трите езика.

Също така се потвърди предположението, че многоезичните онлайн речници, и по-конкретно Glosbe, могат да бъдат използвани като надежден източник на данни при осъществяването на съпоставителни лингвистични изследвания.

Ключови думи: *творителен предикативен, инструментал, translation memory, неанотиран корпус*

Svetlozara Leseva, Ivelina Stoyanova
Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences
E-mail: {zarka| iva}@dcl.bas.bg

Ка класификацији предиката активности који означавају промену

Основни циљ овог рада је да понуди оквир за опис семантичких својстава предиката активности (динамичких предиката) који означавају промену, са примарним фокусом на својство које трпи промену. У овом раду активност се схвата у најширем могућем смислу и укључује различите класе динамичких предиката.

Опис глагола промене заснива се на неколико кључних семантичких карактеристика и обухвата лексичко значење глагола и семантичку класу којој припада, тип промене (узрочна или инхоативна), карактеристике промене (квантна или не, постепена или тренутна), као и релевантне елементе семантичког оквира у *FrameNet-y* (Baker and Fellbaum 2009; Ruppenhofer *et al.* 2016). Студија се базира на семантичким класификацијама глагола које су претходно чврсто успостављене у принстонском *Wordnet-y* (Miller 1995), као и у бугарском *Wordnet*-у (Koeva 2021).

На основу издвојених семантичких карактеристика биће понуђена и тзв. „плитка" класификација глагола промене, са циљем да се обухвате бројне различитости унутар класе и предложи адекватан приступ моделирању њиховог синтаксичког понашања. Класификација је заснована на оквирима из *FrameNet*-а, који групишу глаголе на основу сличних концептуалних својстава и сличног синтаксичког понашања. Сматра се да семантичка својства предиката у великој мери одређују његову синтаксичку реализацију тако што евоцирају одређени семантички оквир, те конфигурацију елемената унутар оквира, као и модус њихове (морфосинтаксичке) експресије.

Наши налази засновани су на аутоматски издвојеним и мануелно одабраним илустративним примерима из бугарског и енглеског језика. Примери и анотација за енглески позајмљени су из корпуса *FrameNet*, док су бугарски ручно анотирани. За оба језика подаци су допуњени примерима из других корпуса, уколико је то било потребно.

Анотирани пример:

FRAME: Cause_expansion; Default: [causal gradual change]
[Universitetat]$_{\text{AGENT}}$ postoyanno **razshiryava** [uchastieto si v kulturniya zhivot]$_{\text{ITEM}}$.
[The University]$_{\text{AGENT}}$ constantly **expands** [its participation in the cultural life]$_{\text{ITEM}}$.

FRAME: Cause_temperature_change; Default: [inchoative gradual change]
[Plamnaloto i litse]$_{\text{ITEM}}$ **se ohladi** [ot ledenata voda]$_{\text{CAUSE}}$.
[Her flushed face]$_{\text{ITEM}}$ **cooled** [by the chill water]$_{\text{CAUSE}}$.

У раду су истражени универзални аспекти концептуалног знања који омогућавају пренос семантичких и синтаксичких информација унутар различитих језика и ресурса. Конфигурација елемената оквира који одређује понашање глагола (евоцирано одређеним оквиром) независна је од језика, као што су то и семантичка ограничења која детрминишу њихову селекцију.

Констелације елемената оквира које добијају синтаксичку експресију међусобним комбиновањем елемената (тзв. „валентни обрасци”, у терминима *FrameNet*-а) махом су валидне и важе у различитим језицима, као што је потврђују подаци за енглески и бугарски. Такође постоји и јасна кореспонденција између синтаксичких категорија и синтаксичких функција елемената унутар семантичких оквира у ова два језика.

Уз то, на основу емпиријске грађе из корпуса анализирана су и специфична својства, као и разлике у синтаксичком и семантичком опису између енглеског и бугарског. Размотрени су случајеви у којима се у ова два језика на различит начин, или на различитим синтаксичким позицијама реализују одређени елементи семантичког оквира. Добијени налази могу бити од значаја и за друге словенске језике који показују уочене граматичке посебности.

Кључне речи: *глаголска семантика, семантика оквира, аспектуалне класе, предикати активности, глаголи промене*

Литература

[1] Colin F. Baker and Christiane Fellbaum. 2009. WordNet and FrameNet as Complementary Resources for Annotation. In Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP ’09), Association for Computational Linguistics, Stroudsburg, PA, USA, pages 125–129.

[2] Svetla Koeva. 2021. The Bulgarian WordNet: Structure and specific features. Papers of Bulgarian Academy of Sciences, 8(1):47–70.

[3] George A. Miller. 1995. WordNet: A lexical database for English. Commun. ACM, 38(11):39–41.

[4] Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. FrameNet II: extended theory and practice. International Computer Science Institute, Berkeley, California.

Maja Matijević
Institut za hrvatski jezik
E-mail: mmatijevic@ihjj.hr

Od rječnika i korpusa do hiponimije i meronimije

Hiponimija i meronimija hijerarhijski su leksičko-semantički odnosi, tj. odnosi koji ustrojavaju leksičke hijerarhije. Hiponimija je leksičko-semantički odnos u kojemu jedinica koja se nalazi na višoj razini u hijerarhiji (hiperonim) označuje vrstu, a jedinice na nižoj hijerarhijskoj razini (hiponimi) označuju predstavnike vrste. Hiperonim je tako primjerice *cvijet*, a njegovi hiponimi *ruža*, *tulipan*, *zumbul* i dr. Meronimija je s druge strane leksičko-semantički odnos u kojemu jedinica koja se nalazi na višoj razini (holonim) označuje cjelinu, a jedinice na nižoj razini (meronimi) označuju dijelove te cjeline, pa je primjerice *tijelo* holonim, a *ruka* i *noga* njegovi meronimi. Pri definiranju hiponima i meronima najčešće se upotrebljavaju ustaljeni obrasci, a leksikografske definicije kao jasno strukturirani i usklađeni dijelovi rječničkoga članka dobar su izvor za uočavanje takvih obrazaca.

U radu će se prikazati kako se analizom rječničkih definicija (točnije, definicija iz prvoga hrvatskog mrežnog rječnika *Mrežnika*) dolazi do leksičko-sintaktičkih obrazaca koji upućuju na hiponimiju i meronimiju te na njihove podvrste (za meronimiju primjerice odnosi

objekt > funkcionalna sastavnica, skup > jedinka, mješavina > sastojak i sl.). Prikazat će se kako se dobiveni obrasci potom propituju u korpusu hrvatskoga jezika (*hrWaC*) i zatim u korpusu specijaliziranoga znanja (*Hrvatskome jezikoslovnom korpusu*). Rječnička i korpusna građa analiziraju se s pomoću Pythona i CQL-a, a osim obrazaca dobivaju se brojni primjeri hiponima i hiperonima te meronima i holonima – što stvara plodno tlo za jezikoslovna istraživanja. Stoga će se također ukratko prikazati glavni zaključci do kojih se dolazi u području leksičke semantike (razlike između hiponimije i meronimije, njihovi dodiri, kvaziodnosi, autohiponimija i automeronimija i sl.).

Istraživanja leksičko-semantičkih odnosa koja se temelje na leksičko-sintaktičkim obrascima češća su u stranome jezikoslovlju, a dobiveni rezultati pokazuju da je takva istraživanja važno prilagoditi flektivnim jezicima kao što je hrvatski (a onda i drugi južnoslavenski jezici) primjerice zbog kategorije padeža, glagolske valencije, prijedložno-padežnih izraza itd.

Ključne reči: *hiponimija, meronimija, leksičko-sintaktički obrazac, rječnik, rječnička definicija, Hrvatski mrežni rječnik – Mrežnik*

Вештачка интелигенција, језички модели и обрада јужнословенских језика

Saša Petalinkar
Univerzitet u Beogradu, Srbija
E-mail: sasa5linkar@gmail.com

Milica Ikonić Nešić
Univerzitet u Beogradu – Filološki fakultet, Srbija
E-mail: milica.ikonic.nesic@fil.bg.ac

Automatizacija kreiranja primera za sinsetove: studija slučaja sa Srpskim Wordnet-om i ChatGPT-om

Wordnet-i su leksički resursi koji organizuju reči prema njihovim značenjima, olakšavajući semantičko razumevanje. Srpski Wordnet, razvijen kao deo projekta Balkanet, predstavlja ključni alat za lingvistička istraživanja i primene. Suštinska komponenta za ljudsko razumevanje sinsetova—grupa sinonimnih reči—jesu ilustrativni primeri. Međutim, pronalaženje primera koji precizno odgovaraju tačnom značenju reči unutar sinseta je dugotrajan i izazovan zadatak. Ovaj rad predlaže inovativno rešenje za ovaj problem korišćenjem jezičkog modela (LM) za generisanje sintetičkih primera za svaku reč u sinsetu, na osnovu datih definicija. Naš pristup koristi sposobnosti naprednih jezičkih modela, posebno ChatGPT-a, za kreiranje kontekstualno prikladnih i semantički tačnih primera. Ova metoda ima za cilj poboljšanje upotrebljivosti i pristupačnosti wordnet-a automatizacijom generisanja ilustrativnih rečenica, čime se štedi značajno vreme i trud za istraživače i leksikografe.

Da bismo ocenili validnost i efikasnost ovog rešenja, generisali smo sintetičke primere za odabrane sinsetove iz Srpskog Wordnet-a i sproveli detaljnu analizu i ocenu tih primera. Kriterijumi evaluacije uključivali su semantičku tačnost, kontekstualnu relevantnost i opštu koherentnost generisanih rečenica.

Naši nalazi ukazuju na to da su primeri generisani pomoću jezičkog modela veoma efikasni u hvatanju nameravanih značenja reči unutar sinsetova. Automatizovani proces ne samo da proizvodi visokokvalitetne primere, već i značajno smanjuje radno-intenzivan zadatak ručnog kreiranja primera. Rezultati pokazuju da jezički modeli poput ChatGPT-a mogu služiti kao snažni alati za poboljšanje leksičkih resursa kao što su wordnet-i.

Ovo istraživanje doprinosi oblasti računske lingvistike predstavljajući skalabilnu i efikasnu metodu za generisanje primera u wordnet-ima. Naglašava potencijal jezičkih modela u podršci razvoju lingvističkih resursa i otvara nove mogućnosti za njihovu primenu u različitim zadacima obrade prirodnog jezika. Budući rad će se fokusirati na dalju optimizaciju procesa generisanja, istraživanje upotrebe drugih naprednih jezičkih modela i širenje primene na druge jezike i leksičke resurse.

U zaključku, integracija jezičkih modela za generisanje sintetičkih primera u wordnet-ima predstavlja značajan napredak u upravljanju leksičkim resursima, nudeći obećavajuće rešenje za izazov kreiranja ilustrativnih primera za sinsetove.

Ključne reči: *jezički model, srpski, sinset, WordNet, ChatGPT*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Milica Ikonić Nešić, Miloš Utvić
Univerzitet u Beogradu – Filološki fakultet, Srbija
E-mail: {milica.ikonic.nesic|milos.utvic}@fil.bg.ac.rs

Tesla-Ner-Nel-Gold skup podataka: studija slučaja na srpsko-engleskom paralelnom korpusu

TESLA-NER-NEL-gold je skup podataka osmišljen kao proširenje postojećeg skupa srpELTeC-gold (Krstev et. al, 2021), namenjenog za treniranje modela za prepoznavanje imenovanih entiteta (NER) i njihovo povezivanje (NEL). Ovaj rad opisuje anotaciju (pod)skupa podataka TESLA-NER-NEL-gold koga čine tekstovi preuzeti iz paralelnog srpsko-engleskog tekstualnog korpusa SrpEngKor (Krstev & Vitas 2011).

Postoje četiri nivoa anotacije u korpusu SrpEngKor-TESLA: vrsta reči (PoS), lema, klasa imenovanih entiteta (NE) i povezivanje imenovanih entiteta sa otvorenom bazom znanja Vikipodaci. Klase imenovanih entiteta obuhvataju demonime (nazive nacionalnosti), imena ličnosti, mesta, organizacija, događaja i umetničkih dela. Takođe, celoviti tekstovi korpusa su povezani sa bibliografskim metapodacima (naslov, autor itd.) i registrom teksta. Statistički opis korpusa SrpEngKor-TESLA biće predstavljen sa detaljima o veličini korpusa (broj tekstova i rečenica), raspodeli PoS i NER klasa, kao i tekstualnih registara.

Anotacija skupa podataka je u prvoj fazi automatska primenom modela SrpCNNeL i Jerteh-355-tesla, u sledećoj fazi anotatori ručno koriguju skup korišćenjem platforme INCEPTION, a završnu ručnu korekciju obavlja supervizor. Povezivanje entiteta se takođe vrši na INCEPTION platformi, gde se prepoznati imenovani entiteti povezuju sa odgovarajućim stavkama Vikipodataka. Pored toga, u radu će biti predstavljen model za prepoznavanje i povezivanje imenovanih entiteta sa vikipodacima treniran na skupu podataka TESLA-NER-NEL-gold.

Osim jednojezične analize anotiranih korpusa, biće ilustrovane prednosti bogatih anotacija na paralelnim korpusima putem upita na različitim jezicima. Ovaj pristup pruža mogućnosti za istraživanje, ekstrakciju i učenje različitih jezičkih obrazaca.

Ključne reči: *paralelni korpusi, imenovani entiteti, NER, NEL, srpski, engleski*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Literatura

[1] Cvetana Krstev and Branislava Šandrih Todorović and Ranka Stanković and Milica Ikonić Nešić. (2021). SrpELTeC-gold - Named Entity Recognition Training Corpus for Serbian. ELG, https://live.european-language-grid.eu/catalogue/corpus/9485, 1.0.

[2] Cvetana Krstev, Duško Vitas, "An Aligned English-Serbian Corpus", In: ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I, Belgrade, 4-6 December 2009, eds. N. Tomović & J. Vujić, pp. 495-508, Faculty of Philology, University of Belgrade, ISBN 978-86-6153-005-0, 2011.

mr Milena Šošić
doktorand na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: milena.sosic@gmail.com, pd202030@alas.matf.bg.ac.rs

prof. dr Ranka Stanković
vanredni profesor na Rudarsko-Geološkom fakultetu Univerziteta u Beogradu
E-mail: ranka.stankovic@rgf.bg.ac.rs

prof. dr Jelena Graovac
vanredni profesor na Matematičkom fakultetu Univerziteta u Beogradu
E-mail: jgraovac@matf.bg.ac.rs

Social-Emo.Sr: Emocionalna višeznačna kategorizacija konverzacionih poruka sa društvenih mreža X i Reddit

U digitalnom okruženju južnoslovenskih jezika, analiza emocija u tekstovima na društvenim mrežama postaje sve važnija za razumevanje javnog mnjenja, kreiranje personalizovanog sadržaja i analizu međusobnih interakcija korisnika. U okviru ovog rada predstavljamo detaljnu metodologiju i rezultate označavanja korpusa na srpskom jeziku prema Plutčikovom modelu kategorizacije, koji prepoznaje osam osnovnih emocionalnih kategorija, kao što su radost, tuga, bes, strah, poverenje, gađenje, iščekivanje i iznenađenje. Cilj istraživanja je da se analizira emocionalni sadržaj tekstova preuzetih sa društvenih mreža X (nekada Twitter) i Reddit, od kojih svaka kolekcija sadrži oko 17,000 pojedinačnih poruka i približno 5,000 kompletnih konverzacija. Proces označavanja korpusa uključivao je nekoliko faza: prikupljanje i pripremu podataka, ručno označavanje od strane stručnih lica, verifikaciju preciznosti označavanja i statističku analizu harmonizovanih oznaka. Korišćenjem pristupa višeznačnog označavanja, omogućena je bogatija i kvalitetnija analiza emocionalnih stanja, sa posebnim značajem na primenu u analizama složenih emocionalnih sadržaja koji se mogu pronaći na društvenim mrežama.

Za prikupljanje podataka korišćeni su automatizovani alati za preuzimanje konverzacija napisanih na srpskom jeziku sa naloga na društvenim mrežama koji obrađuju aktuelne društvene, političke, muzičke i sportske teme. Priprema podataka obuhvatila je dodatnu selekciju poruka da bi se obezbedio kvalitet njihovog sadržaja, uz uslov održanja konverzacione strukture preuzetih podataka. U okviru pripreme podataka, poruke su unapred označene automatskim metodama korišćenjem klasičnih, ali i naprednih tehnika računarske lingvistike, u cilju unapređenja efikasnosti procesa ručnog označavanja. Timovi lingvista i psihologa su automatski dodeljene oznake proveravali i ocenjivali u pogledu njihove verodostojnosti za tekstualni sadržaj kojima su dodeljene. U clju obezbeđivanja visoke tačnosti i konzistentnosti, korišćene su standardizovane procedure za obuku anotatora i proveru njihovih ocena kroz statističke mere za pouzdanost označavanja. Analiza pouzdanosti označavanja je pokazala da je moguće klasifikovati emocije u tekstovima sa društvenih mreža na srpskom jeziku koristeći Plutčikov model. Statistička analiza podataka je otkrila značajne distribucije emocija u porukama i pružila uvid u emocionalne reakcije korisnika na različite emocionalne nadražaje i tematske sadržaje.

Višeznačno kategorizovan emocionalni korpus na srpskom jeziku Social-Emo.SR predstavlja značajan iskorak ka dubljem razumevanju emocionalne dinamike na društvenim mrežama među korisnicima. Osim obogaćivanja lingvističkih resursa za srpski jezik, ovaj korpus otvara nove mogućnosti za primenu u istraživanjima, komercijalnim aplikacijama i

unapređenju analize mentalnog zdravlja populacije. Potencijalna primena savremenih metodologija nad razvijenim korpusom omogućila bi kreiranje korisnih alata za prepoznavanje i reflektovanje složenosti ljudskih emocija u aktuelnom digitalnom svetu na srpskom govornom području. Korpus će biti objavljen pod javnom licencom CC-BY-4.0.

Ključne reči: *emocije, Plutčikov model, označavanje, korpus, društvene mreže, srpski jezik*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Mihailo Škorić
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: mihailo@jerteh.rs

Novi jezički modeli za južnoslovenske jezike

Izlaganje će predstaviti izazove i perspektive modelovanja južnoslovenskih jezika, sa posebnim osvrtom opšte jezičke modele građene na arhitekturi transformera (BERT, GPT), na dostupne skupove tekstova za obučavanje tih modela, te kvantitet i kvalitet tih skupova. Izlaganje će ponuditi pregled dostupnih skupova i modela, dok će posebna pažnja biti posvećena najnovijim korpusima tekstova. Prvi korpus, *Kišobran*, predstavlja krovni veb korpus južnoslovenskih jezika i ujedno trenutno najveći korpus tekstova na našim prostorima koji broji preko osamnaest milijardi reči i uključuje sve ostale trenutno dostupne južnoslovenske veb korpuse. Drugi korpus, *S.T.A.R.S*, na jednom mestu okuplja akademske radove pisane na srpskom jeziku i uključuje pre svega jedanaest hiljada disertacija preuzetih sa platforme NARDUS, ali i veliki broj naučnih i stručnih radova preuzetih iz različitih otvorenih repozitorijuma koji su uvršteni u sistem *eNauka*. Osim toga, biće reči o akademskih korpusima ostalih južnoslovenskih jezika, koji su nastali od radova pohranjenih na različitim veb platformama: DABAR (za hrvatski jezik), repozitorijuma univerziteta u Mariboru, Ljubljani, Primorskoj i Novoj Gorici i repozitorijuma DiRROS i REVIS (za slovenački jezik), repozitorijuma univerziteta u Zenici, Sarajevu i Istočnom Sarajevu (za bosanski jezik), repozitorijuma Univerziteta Goce Delčev i Sv. Kliment Ohridski (za makedonski jezik) i repozitorijuma Univerziteta Crne Gore (za crnogorski). Naposletku, biće reči o novim modelima za vektorizaciju teksta pisanog na južnoslovenskim jezicima, a koji su obučavani korišćenjem upravo navedenih korpusa tekstova. Biće predstavljena analiza njihovih performansi na nekolicini prethodno utvrđenih zadataka sa osvrtom na unapređenja koja su ostvarena u odnosu na rezultate modela obučavanih na prethodnoj generaciji korpusa.

Ključne reči: *Veliki korpusi teksta, jezički modeli, južnoslovenski jezici*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Danka Jokić, Ranka Stanković, Jelena Jaćimović
Društvo za jezičke resurse i tehnologije JeRTeh
E-mail: danka.jokic@afrodita.rcub.bg.ac.rs; ranka.stankovic@rgf.bg.ac.rs; jelena.jacimovic@stomf.bg.ac.rs

Grafovi znanja u doba velikih jezičkih modela: prilike i izazovi

Pojava velikih jezičkih modela (eng. Large Language Models ili LLMs) je značajno uticala na oblast veštačke inteligencije, naročito u oblastima obrade prirodnog jezika i generisanju teksta. Međutim, ključno ograničenje ovih modela leži u nedostatku strukturiranog znanja i sposobnosti zaključivanja, što otežava njihovu primenu u stvarnom svetu, gde se zahteva tačnost iznetih činjenica i zaključivanje na osnovu konteksta. S druge strane, grafovi znanja nude primamljivo rešenje. Oni pružaju bogat izvor strukturiranog znanja, tako što predstavljaju entitete i njihove relacije u mašinski čitljivom formatu. Ova komplementarnost pruža jedinstvenu priliku za istraživanje njihovog simbiotskog odnosa: primenu grafova znanja kako bi se veliki jezički modeli osposobili za razvoj veštačke inteligencije sledeće generacije.

Sinergija sa grafovima znanja pruža mogućnosti značajnog unapređenja velikih jezičkih modela u nekoliko ključnih oblasti. Najpre, postavljanje strukturiranog znanja iz grafova znanja u temelje velikih jezičkih modela, može značajno poboljšati njihovu sposobnost da razumeju činjenične informacije i generišu tačnije i pouzdanije odgovore na složena pitanja. Pored toga, grafovi znanja pružaju neophodan kontekst i relacije među pojmovima, što bi omogućilo velikim jezičkim modelima da obavljaju sofisticiranije rezonovanje i ugrade zdravorazumsko znanje u svoje procese rasuđivanja do nivoa razumevanja koji je sličniji ljudskom. Štaviše, eksplicitne relacije unutar grafova znanja mogu se iskoristiti da se objasni rezonovanje koje stoji iza izlaza iz velikih jezičkih modela, čime se direktno rešava ključni izazov u objašnjivoj veštačkoj inteligenciji.

Iako je potencijal neosporno veliki, izazovi sinergije velikih jezičkih modela i grafova znanja takođe zahtevaju pažnju. Razvoj efikasnih metoda, koji bi omogućili da veliki jezički modeli istovremeno uče iz tekstualnih podataka i grafova znanja, je ključno za uspešnu integraciju. Takođe, obezbeđivanje konzistentnosti i kvaliteta podataka u grafovima znanja je od suštinskog značaja, jer nepotpune ili netačne informacije mogu dovesti do pristrasnih ili pogrešnih rezultata velikih jezičkih modela.

Uprkos navedenim izazovima, integracija velikih jezičkih modela i grafova znanja ima ogroman potencijal da reformiše različite aplikacije bazirane na veštačkoj inteligenciji. Veliki jezički modeli potpomognuti grafovima znanja mogu da pruže tačnije i sveobuhvatnije odgovore u sistemima za odgovaranje na pitanja. Inteligentni asistenti integrisani sa grafovima znanja mogu razumeti i odgovoriti na zahteve korisnika sa više konteksta i na osnovu činjenica. Pored toga, sinergija velikih jezičkih modela i grafova znanja može dovesti do generisanja prirodnog jezika, koje je tačnije i bazirano na relevantnom kontekstu, kao i do stvaranja sofisticiranijih i dinamičnijih sistema za predstavljanje znanja.

Istraživanje sinergije između velikih jezičkih modela i grafova znanja ključno je za unapređenje veštačke inteligencije. Rešavanjem izazova i traženjem efikasnih metoda integracije, možemo utrti put za razvoj sledeće generacije aplikacija zasnovanih na veštačkoj inteligenciji, koje karakterišu poboljšano razumevanje, rezonovanje i sposobnosti predstavljanja znanja. U ovom radu ćemo pokazati kako grafovi znanja mogu unaprediti sposobnosti rezonovanja velikih jezičkih modela u pogledu bezbednosti i moderacije štetnog sadržaja na internetu u tekstovima na srpskom jeziku.

Ključne reči: *grafovi znanja, veliki jezički modeli, obrada prirodnog jezika, strukturirano znanje, kvalitet podataka, objašnjiva veštačka inteligencija, bezbednost sadržaja na internetu*

Zahvalnica: *Ovo istraživanje je podržao Fond za nauku Republike Srbije, #7276, Text Embeddings - Serbian Language Applications - TESLA.*

Literatura

[1] Agrawal, G., Kumarage, T., Alghamdi, Z., & Liu, H. (2024). Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 3947–3960). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.219.

[2] Kau, A., He, X., Nambissan, A., Astudillo, A., Yin, H., & Aryani, A. (2024). Combining Knowledge Graphs and Large Language Models (arXiv:2407.06564). arXiv. https://doi.org/10.48550/arXiv.2407.06564.

[3] Lin, J. (2022). Leveraging World Knowledge in Implicit Hate Speech Detection. In L. Biester, D. Demszky, Z. Jin, M. Sachan, J. Tetreault, S. Wilson, L. Xiao, & J. Zhao (Eds.), Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI) (pp. 31–39). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.nlp4pi-1.4.

[4] Pan, J., Razniewski, S., Kalo, J. C., Singhania, S., Chen, J., Dietze, S., ... & Graux, D. (2023). Large Language Models and Knowledge Graphs: Opportunities and Challenges. Transactions on Graph Data and Knowledge.

[5] Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2024.3352100.

Никола Јанковић
Филолошки факултет, Универзитет у Београду
E-mail: nikolajankovickv@gmail.com

Јована Иваниш
Институт за српски језик САНУ
E-mail: jovana.ivanis@gmail.com

Употреба модела Whisper Large V3 Sr за транскрипцију говора на српском језику у програмском језику *Пајтон* на платформи *ГУГЛ КОЛАБ*

У овом раду представљен је скрипт у програмском језику *Пајтон* (енгл. *Python*) на платформи *Гугл колаб* (енгл. *Google Colab*) који користи фино подешени модел за транскрипцију говора на српском језику *Whisper Large v3 Sr* (https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb), чиме се омогућава бесплатно, квалитетно и једноставно транскрибовање говора на српском језику у текст. Мотивација за креирање овог скрипта проистекла је из недостатка доступних алата који истраживачима пружају једноставно коришћење овог модела без потребе за напредним

техничким знањима и значајним рачунарским ресурсима. Овај скрипт омогућује једноставно учитавање аудио-фајлова, њихову транскрипцију помоћу модела *Whisper Large v3 Sr* и преузимање транскрипције фајлова у текстуалном формату. У раду ће најпре укратко бити описан горенаведени модел који скрипт користи, а затим ће детаљно бити описан начин функционисања самог скрипта, уз упутство за коришћење. У раду ће такође бити представљени проблеми које је било неопходно превазићи за успешну примену овог приступа, као што је потреба за аутоматском сегментацијом аудио-фајлова, одређивање оптималних параметара сегментације и омогућавање подршке за различите формате аудио-фајлова. Осим тога, биће представљено неколико приступа везаних за смањивање броја грешака у транскрипцији. На крају, биће приказана евалуација тачности и ефикасности транскрипције, као и статистичка анализа уочених грешака. Сматрамо да овај алат може бити од великог значаја за истраживаче, с обзиром на то да убрзава процес обраде говорних података и омогућава обраду велике количине аудио-материјала у кратком периоду и пружа конзистентан и поновљив метод транскрипције, што је нарочито значајно за научну методологију и поновљивост истраживања везаних за језик. Поред тога, сматрамо да алат може бити користан и другим корисницима, с обзиром на то да омогућава једноставно креирање титлова за видео-садржаје, претварање аудио-бележака у текстуални формат, стварање аутоматски генерисаних транскрипата за особе са оштећеним слухом, као и стварање текстуалних архива аудио-садржаја.

Кључне речи: *транскрипција говора, српски језик, Пајтон, Whisper Large v3, Гугл колаб, NLP*

Ana Kovačević
Fakultet bezbednosti, Univerzitet u Beograd
E-mail: kana@fb.bg.ac.rs

Lažne vesti i generativna veštačka inteligencija: rizici i moguća rešenja

Fenomen lažnih vesti je prisutan od samih početaka ljudske komunikacije, a ponovljeno izlaganje lažnim informacijama često dovodi do njihovog prihvatanja kao istinitih. Međutim, u savremenom društvu, lažne vesti postaju izuzetno opasne zbog njihovog jednostavnog, jeftinog i uverljivog načina kreiranja pomoću generativne veštačke inteligencije, posebno velikih jezičkih modela. Ovi modeli, koji su doživeli značajan napredak, omogućavaju generisanje velike količine sadržaja koji zvuči verodostojno i na srpskom jeziku. Pored toga, omogućeno je kreiranje personalizovanog sadržaja specifičnog za određene grupe ili pojedince, uz navođenje navodno kredibilnih izvora, što dodatno pojačava uticaj lažnih vesti. U svetu preopterećenom informacijama, ovaj problem postaje još izraženiji. Jedan od ključnih kanala širenja lažnih vesti su društvene mreže, čiji je primarni cilj zadržavanje korisničke pažnje sadržajem koji potvrđuje njihova uverenja. Ovaj proces rezultira efektom eho komore, gde korisnici dobijaju informacije koje potvrđuju njihova prethodna uverenja, čime se otežava prepoznavanje lažnih informacija. Sekundarni problem lažnih vesti je izazivanje nepoverenja i konfuzije među korisnicima, zbog čega dolazi do gubitka poverenja i u tačne informacije. Velika količina podataka koja postoji na društvenim mrežama naglašava potrebu za primenom veštačke inteligencije, ne samo za prikazivanje relevantnog sadržaja, već i za detekciju lažnih vesti. Pristupi zasnovani na veštačkoj inteligenciji mogu se koristiti za prepoznavanje lažnih informacija, ali se postavlja pitanje koliko su algoritmi zaista transparentni i pouzdani u tom zadatku. U radu će biti

predstavljene mogućnosti primene veštačke inteligencije u kreiranju lažnih vesti, kao i moguća rešenja ovog problema.

Ključne reči: *veštačka inteligencija, veliki jezički modeli, lažne vesti, rizici veštačke inteligencije*

Zahvalnica: *Rad je nastao u okviru projekta koji finansira Fond za nauku Republike Srbije u okviru Programa "ИДЕЈЕ" - Management of New Security Risks - Research and Simulation Development, NEWSIMR&D, #7749151*

Maram Alharbi
Lancaster University, Jazan University
E-mail: m.i.alharbi@lancaster.ac.uk

Ruslan Mitkov
Lancaster University and University of Alicante
E-mail: r.mitkov@lancaster.ac.uk

Poređenje pristupa zasnovanog na pravilima i dubokog učenja za razumevanje osećanja

U domenu analize zadovoljstva ugostiteljskim uslugama koja se brzo razvija, tačno razumevanje osećanja i mišljenja korisnika iz recenzija hotela je sve važnije za inteligentno upravljanje tržištem i poboljšanje kvaliteta usluge. Ovo istraživanje se fokusira na kategorizaciju recenzija hotela na pozitivne i negativne komentare koristeći niz modela analize osećanja. Ispitujemo pristupe zasnovane na pravilima—kao što su VADER, AFINN i TektBlob—, kao i metode dubokog učenja, uključujući model T5 transformatora i BiLSTM. Dok modeli zasnovani na pravilima koriste unapred definisane leksikone osećanja, nudeći jednostavnost i interpretabilnost, oni često ne uspevaju da uhvate složenosti emocionalnih nijansi i osećanja zavisnih od konteksta. Nasuprot tome, modeli dubokog učenja, posebno oni koji koriste napredne transformerske modele, pružaju sofisticiranije razumevanje jezičkih kompleksnosti, omogućavajući efikasnije otkrivanje osećanja, čak i u složenim kontekstima. Naši eksperimenti pokazuju da modeli zasnovani na pravilima postižu umerene performanse, sa F1 rezultatima od 0,79 za TektBlob, 0,77 za AFINN i 0,78 za VADER. Iako efikasne, ove metode se često bore sa kontekstualnim suptilnostima osećanja. Nasuprot tome, modeli dubokog učenja pokazali su superiorne performanse, pri čemu je model transformatora T5 postigao F1 rezultat od 0,96, a BiLSTM model je dostigao 0,93. Ovo naglašava potencijal tehnika dubokog učenja za analizu sentimenta u recenzijama kupaca, nudeći precizniju i nijansiraniju klasifikaciju osećanja od tradicionalnih metoda. Upoređujemo ove pristupe u dva skupa podataka recenzija hotela, procenjujući uticaj različitih tehnika prethodne obrade i modela analize sentimenta u smislu tačnosti, preciznosti i odzivu. Naši nalazi naglašavaju superiorne performanse modela dubokog učenja, posebno transformera T5, u preciznoj klasifikaciji osećanja i rešavanju izazova koje predstavljaju mešana osećanja u komentarima. Bez obzira na to, modeli zasnovani na pravilima zadržavaju svoju korisnost u scenarijima gde je računarska efikasnost prednost. Ovo istraživanje pruža sveobuhvatnu procenu metoda analize sentimenta u hotrlijerstvu, nudeći uvide koji mogu poboljšati i korisničko iskustvo i poslovnu strategiju. Upoređivanjem tradicionalnih i naprednih tehnika analize osećanja, doprinosimo dubljem razumevanju performansi modela i njihove praktične primenljivosti u realnim okruženjima.

Ključne reči: *Analiza osećanja, pristup zasnovan na pravilima, duboko učenje, transformeri, BiLSTM, T5, VADER, TextBlob, AFINN*

Literatura

[1] R. Mitkov, The Oxford Handbook of Computational Linguistics 2nd edition, Oxford University Press, 2022.

[2] E. Demir, M. Bilgin, Sentiment analysis from Turkish news texts with Bert-based language models and machine learning algorithms, in: 2023 8th International Conference on Computer Science and Engineering (UBMK), IEEE, 2023, pp. 01–04.

[3] A. Ameur, S. Hamdi, S. Ben Yahia, Sentiment analysis for hotel reviews: A systematic literature review, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3605152. doi:10.1145/3605152.

[4] S. Mutmainah, D. H. Fudholi, Leveraging Bilstm and Lda for analyzing and dashboarding user feedback in applications, JURNAL MEDIA INFORMATIKA BUDIDARMA 8 (2024) 51–61.

[5] C. Hutto, E. Gilbert, Vader: A parsimonious rule- based model for sentiment analysis of social media text, in: Proceedings of the international AAAI conference on web and social media, volume 8, 2014, pp. 216–225.

[6] J.-P. Colson, Multi-word units in machine translation: why the tip of the iceberg remains problematic–and a tentative corpus-driven solu- tion, Computational and Corpus-based Phraseology (2019) 145.

[7] Mitkov, Ruslan, Computer vs. human intelligence. keynote speech at the Refinitiv conference, City of London., 2019.

[8] R. Stuckardt, Machine-learning-based vs. manually designed approaches to anaphor resolution the best of two worlds. proceedings of the dis- course anaphora and anaphora resolution colloquium, daarc'4, 211-216. Lisbon, Portugal., 2002.

[9] R. Stuckardt, Three algorithms for competence- oriented anaphor resolution. proceedings of the discourse anaphora and anaphora resolution colloquium, daarc'5, 157-163. Sao Miguel, Portugal., 2003.

[10] R.Stuckardt, A machine learning approach to preference strategies for anaphor resolution. in Antonio Branco, Tony McEnery, and Ruslan Mitkov (eds.), anaphora processing: Linguistic, cognitive and computational modelling, 47-72. John Benjamins, Amsterdam/ Philadelphia, 2004.

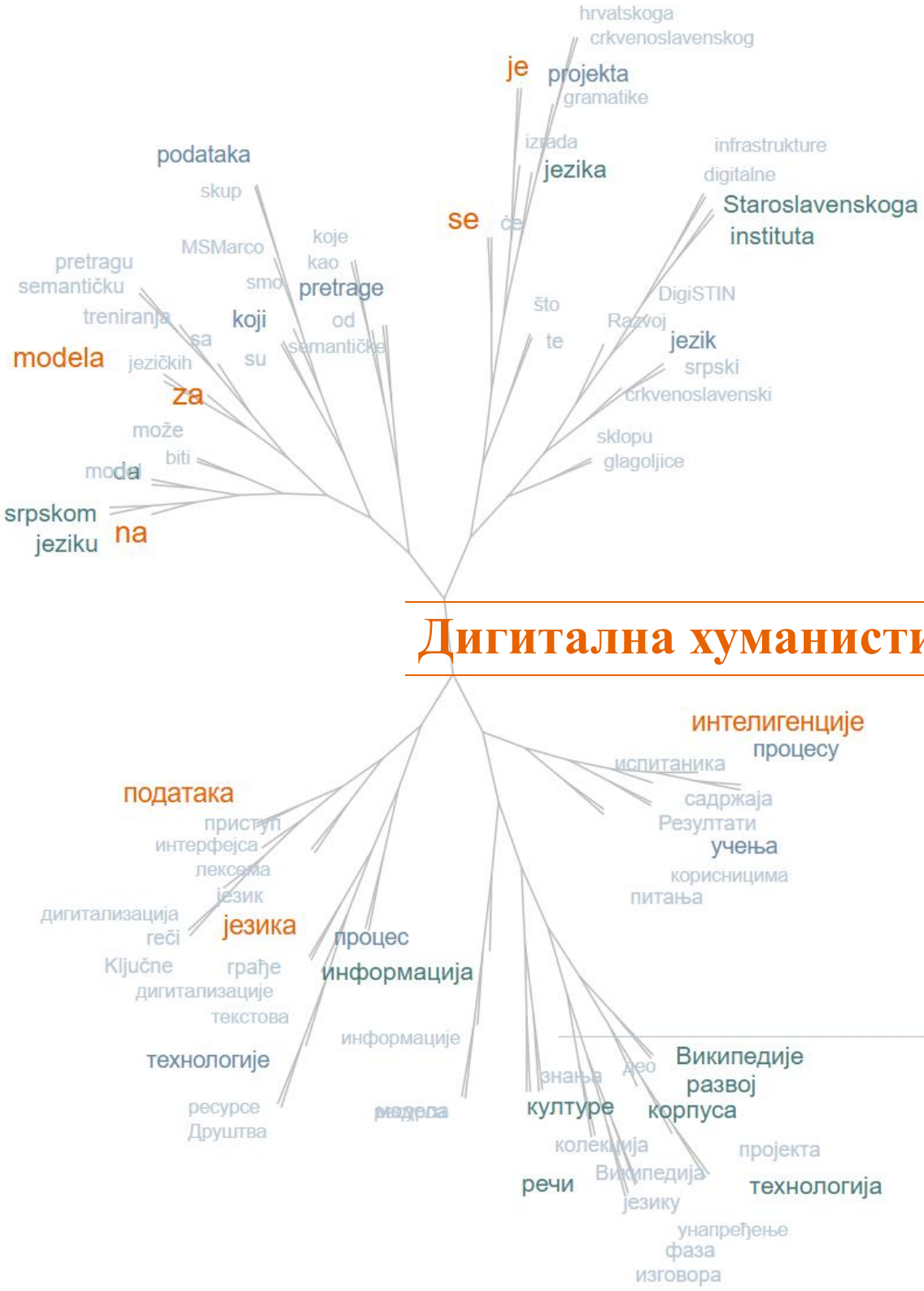
[11] G. Sreenivas, K. M. Murthy, K. Prit Gopali, N. Eedula, M. H R, Sentiment analysis of hotel reviews - a comparative study, in: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), 2023, pp. 1–9. doi:10.1109/I2CT57861.2023.10126445.

[12] M.R. Orasan, C., Recent Developments in Natural Language Processing. In R. Mitkov (Ed.), The Ox- ford Handbook of Computational Linguistics 2nd edition. Oxford University Press., Oxford Univer- sity Press, 2021.

[13] C. Raffel, N.Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, CoRR abs/1910.10683 (2019). URL: http: //arxiv.org/abs/1910.10683. arXiv:1910.10683.

[14] K. Pipalia, R. Bhadja, M. Shukla, Comparative analysis of different transformer based architectures used in sentiment analysis, in: 2020 9th International Conference System Modelling and Advancement in Research Trends (SMART), 2020, pp. 411–415. doi:10.1109/SMART50582.2020.9337081.

[15] A. Pathak, et al., Comparative analysis of trans- former based language models, in: CS & IT Conference Proceedings, volume 11, CS & IT Conference Proceedings, 2021.



Дигитална хуманистика

Милена М. Стојановић
Филолошки факултет Универзитета у Београду
E-mail: stojanovic.milena77@gmail.com

Значења и положај појединих лексема из сфере вештачке интелигенције у лексичком фонду српског језика и перцепција нових технологија – изазови модерног времена

У овом истраживању анализиране су из угла семантике и лингвокултуролошки четири лексеме из сфере вештачке интелигенције (*вебинар, трол, фишинг* и *четовање*) које нису део *Речника српског језика* Матице српске. Желело се дознати какав став имају студенти и средњошколци према новим технологијама и у којој мери познају семантику наведених лексема. Оправдање за лингвокултуролошки приступ теми виђен је у чињеници да суштину језика чини субјективни поглед на свет, што је полазиште ових истраживања. Семантички приступ виђен је у чињеници да се значење лексема често надограђује у одређеном контексту те да је терминологизација процес интелектуализације језика, а детерминологизација вид демократизације језика. Путем онлајн упитника, подељеног у четири целине, добијени су резулати који су интерпретативном анализом, помоћу систематизације, дескрипције и тумачења показали какви су ставови испитаника у вези са темом. Структурно питања су се тицала значаја вештачке интелигенције у процесу учења, колико су корисницима вештачке интелигенције позната значења термина са којима се свакодневно сусрећу у дигиталном окружењу, постоје ли предрасуде и стереотипи о коришћењу нових технологија у животу и у настави, а било је и питања у вези са ставовима о језику, тј. кохабитацији језика и вештачке интелигенције. Упитник је попунило укупно 173 испитаника, од којих је било 72,3% жена и 27,7% мушкараца. Највише испитаника било је из Београда 132 (76,3%), а затим из Војводине 12 (6,9%), Западне Србије 11 (6,4%), а намање из Источне Србије 3 (1,7%). Процентуално, највише испитаника чинила је средњошколска популација – 63%, а студената је било 38,2%. Резултати су показали да семантика анализираних лексема показује висок ниво компетенције говорника, али залази и у домен прагматике и лингвокултурологије. Анализа је потврдила да терминологизација јесте процес интелектуализације језика, а детерминологизација процес демократизације те да се на тај начин лексички фонд значајно богати. Резултати улоге вештачке интелигенције у процесу учења показују да популација која стиче знања има поларизован став према употреби вештачке интелигенције у процесу учења, а изричити су били да ChatGPT, као додатно средство традиционалној настави, није добра идеја (*не* = 73, а *да* = 45, *не знам* = 57). Резултати показју да се савременим технологијама највише замера потенцијална злоупотреба и често упитан квалитет пласираних информација те да професори не би, употребом модерних технологија, побољшали квалитет наставе.

Кључне речи: *вешташка интелигенција, ChatGPT, модерне технологије, лингвокултурологија, стереотип, предрасуде, дигитално окружење*

Небојша Ратковић
Викимедија Србије, Менаџер образовног програма
E-mail: nebojsa.ratkovic@vikimedija.org

Интеграција Википедије на српском језику у образовне системе и унапређење језичких технологија

Википедија на српском језику представља један од најобимнијих и најдоступнијих извора знања у дигиталној ери. Интеграција Википедије на српском језику у образовне системе може значајно допринети модернизацији образовања, побољшању језичке и технолошке писмености, и подршци развоју језичких технологија међу ученицима и студентима. Овај рад истражује могућности и изазове повезане са применом Википедије у наставним процесима, са посебним нагласком на изградњу дигиталних корпуса и унапређење језичких технологија, док истовремено разматра стратешке кораке за интеграцију Википедије у образовне системе на националном нивоу, уз препоруке за политику образовања која би омогућила систематичну употребу Википедије у школама и универзитетима, укључујући обуку наставника, развој наставних планова и креирање дигиталних платформи за подршку овој интеграцији.

Први део рада анализира улогу Википедије као наставног алата и њен потенцијал у обогаћивању наставних материјала и методологија. Осим тога, биће представљено како Википедија може послужити као платформа за развој критичког мишљења, истраживачких вештина и дигиталне писмености ученика. У ту сврху, биће приказани случаји из школа и универзитета где је Википедија успешно интегрисана у наставу, као и повратне информације наставника и ученика о овој интеграцији.

Други део рада фокусира се на техничке аспекте коришћења Википедије у образовању. Размотриће се како се чланци са Википедије могу користити за изградњу дигиталних корпуса који служе као основа за развој и унапређење језичких технологија. Примена Википедије у образовању значајно доприноси изградњи дигиталних корпуса кроз обогаћивање и стандардизацију језичког садржаја, што омогућава бољу анализу и развој језичких технологија, као што су обрада природног језика и алати за аутоматско превођење. Овај процес подстиче систематичну дигитализацију језика и стварање ресурса који су од кључног значаја за даље унапређење и примену језичких технологија у различитим областима.

Кључне речи: *Википедија, образовање, дигитални корпус, језичке технологије*

Ana Mihaljević
Staroslavenski institut, Zagreb, Hrvatska
E-mail: amihaljevic@stin.hr

Hrvatski crkvenoslavenski jezik i glagoljica u digitalnome okružju

Od 1. siječnja 2024. godine na Staroslavenskome se institutu u Zagrebu kao jedan od NextGenerationEU projekata, koje financira Europska unija, provodi projekt *Razvoj modela digitalne infrastrukture Staroslavenskoga instituta - DigiSTIN*. Cilj je projekta osmisliti model razvoja digitalne infrastrukture Staroslavenskoga instituta, koji će se primijeniti u praksi.

Већина је пројекта усмјерена на развој институтских мрежних страница *stin.hr*, за које се настоји да, осим што доносе информације о организацији и дјеловању Старославенскога института, буду и средиште информирања о хрватскоме црквенославенском језику те хрватској глагољици и глагољаштву у цјелини.

Један је од главних циљева пројекта израда мрежне иначице *Рјечника црквенославенскога језика хрватске редакције*, који је досад био објављиван само у тиску, а чија је израда темељна дјелатност Старославенскога института. У излагању ће се представити тренутачно стање његове дигитализираности. Упоредно с дигитализацијом Рјечника дигитализира се и грађа за његову израду (раније су дигитализирани мјесни и азбучни каталог, а сад се дигитализирају упоредни хрватски црквенославенски - латински - грчки геслари) те се израђује претраживи одостражни рјечник. У тијеку је и израда мале е-граматике хрватскога црквенославенскога језика на темељу институтске тискане граматике *Хрватски црквенославенски језик*.

У склопу пројекта DigiSTIN планира се и проширивање корпуса хрватскога црквенославенскога језика доступнога у бази *beram.stin.hr*. У тијеку је скенирање и очитавање (OCR и HTR) рукописа и тискања извора. Како би се поболјшала квалитета трансліитерација и брзина процеса, развијају се нови модели за очитавање рукописа у склопу платформе Transkribus, прилагођени за поједине специфичне облике глагољице.

У склопу мрежних страница Старославенскога института отворене су подстранице *Глагољица у школи*, намијењена за популаризацију глагољице у школскоме узрасту, на којој се објављују садржаји намијењени уporabi на настави Хрватскога језика и других предмета те подстраница с игрицама за усвајање глагољице, али и лексика и граматике хрватскога црквенославенскога језика. У излагању ће се представити полазне точке и циљеви пројекта, али и оно што је досад направљено те планови за сљедеће раздоблје.

Кључне речи: *glagoljica, hrvatski crkvenoslavenski jezik, Razvoj modela digitalne infrastrukture Staroslavenskoga instituta - DigiSTIN, mrežni rječnik, obrazovne igre, korpus, popularizacija*

Miloš Košprdić, Gorana Gojić, Adela Ljajić, Dragiša Mišković
Istraživačko-razvojni institut za veštačku inteligenciju Srbije
E-mail: {milos.kosprdic|gorana.gojic|adela.ljajic|dragisa.miskovic}@ivi.ac.rs

Razvoj modela semantičke pretrage za srpski jezik

Razvoj velikih jezičkih modela predstavlja značajan napredak u oblasti obrade prirodnih jezika, omogućavajući efikasnu semantičku pretragu i razumevanje teksta. U ovom radu predstavljamo proces treniranja velikog jezičkog modela za semantičku pretragu na srpskom jeziku, fokusirajući se na zadatak rangiranja pasusa (engl. passage ranking). Model koji koristimo je zasnovan na msmaarco-bert-base-dot-v5 arhitekturi i prilagođen za asimetričnu semantičku pretragu.

Da bismo omogućili treniranje modela na srpskom jeziku, koristili smo MSMarco skup podataka koji smo automatski preveli sa engleskog na srpski koristeći Google prevodilac. Ovaj skup podataka obuhvata širok spektar pitanja i odgovora, omogućavajući modelu da nauči bogatstvo semantičkih veza na srpskom jeziku.

Cilj ovog rada je da prikaže dosadašnje rezultate u formiranju modela semantičke pretrage na srpskom jeziku i demonstrira mogućnost uspešnog treniranja modela za semantičku

pretragu na jeziku sa ograničenim resursima. Pored tehničkih aspekata treniranja, predstavljeni su i izazovi koji su se javili tokom procesa prevođenja podataka, kao i strategije koje smo primenili za njihovo prevazilaženje.

Proces treniranja modela obavljen je kroz 60 epoha sa veličinom batch-a od 64, koristeći jedan GPU A100 od 40GB na Nacionalnoj platformi za veštačku inteligenciju u Kragujevcu. Model je testiran na delu MSMarco podskupa za testiranje upita i odgovora, gde je postigao zadovoljavajuće performanse u pogledu tačnosti i relevantnosti pretrage, što se može videti iz Tabele

Rezultati evaluacije modela na engleskom i srpskom jeziku

model	**Acc@10**	**P@10**	**R@10**	**MRR@10**	**NDCG@10**	**MAP@100**
engleski	68.51	6.91	68.10	0.3700	0.4435	0.3802
srpski	54.03	5.44	53.7	0.2923	0.3501	0.3024

Acc@10 – Accuracy, P@10 – Precision, R@10 – Recall, MRR@10 – Mean Reciprocal Rank, NDGC@10 – Normalized Discounted Cumulative Gain, MAP@100 – Mean Average Precision.

Naši rezultati pokazuju da model može zadovoljavajuće da rangira pasuse na srpskom jeziku, demonstrirajući robustnost i prilagodljivost arhitekture msmaarco-bert-base-dot-v5. Ovaj model doprinosi razvoju NLP alata za srpski jezik, otvarajući nove mogućnosti za primenu semantičke pretrage u raznim domenima.

Rad doprinosi razvoju jezičkih modela za manje zastupljene jezike, pružajući okvir i metodologiju koja može biti primenjena i na druge jezike sa sličnim izazovima u pretprocesiranju i treniranju. Naši rezultati potvrđuju da automatsko prevođenje podataka uz određene korekcije može biti efikasan pristup za treniranje visokokvalitetnih jezičkih modela, čime se omogućava bolja podrška za srpski jezik u digitalnom okruženju. Dalji rad će biti usmeren ka unapređenju modela kroz dodatno fino podešavanje i evaluaciju na specifičnim zadacima semantičke pretrage, kao i na primeni u realnim sistemima pretrage.

Ključne reči: *semantička pretraga, zadatak rangiranja pasusa, MSMarco skup podataka, veliki jezički modeli, srpski jezik*

Др Снежана Петровић
Институт за српски језик САНУ
E-mail: snezzanaa@gmail.com

Др Мирјана Петровић-Савић
Институт за српски језик САНУ
E-mail: mirjana.petrovic@isj.sanu.ac.rs

Др Ана Шпановић
Институт за српски језик САНУ
E-mail: tesicana@gmail.com

Мср Ленка Бајчетић
Иновациони центар Електротехничког факултета у Београду д.о.о.
E-mail: lenka.bajcetic@gmail.com

Мср Матија Нешовић
Институт за српски језик САНУ
E-mail: nesovic1998@gmail.com

Мср Јована Тодорић
Институт за српски језик САНУ
E-mail: jovanatodoric080@gmail.com

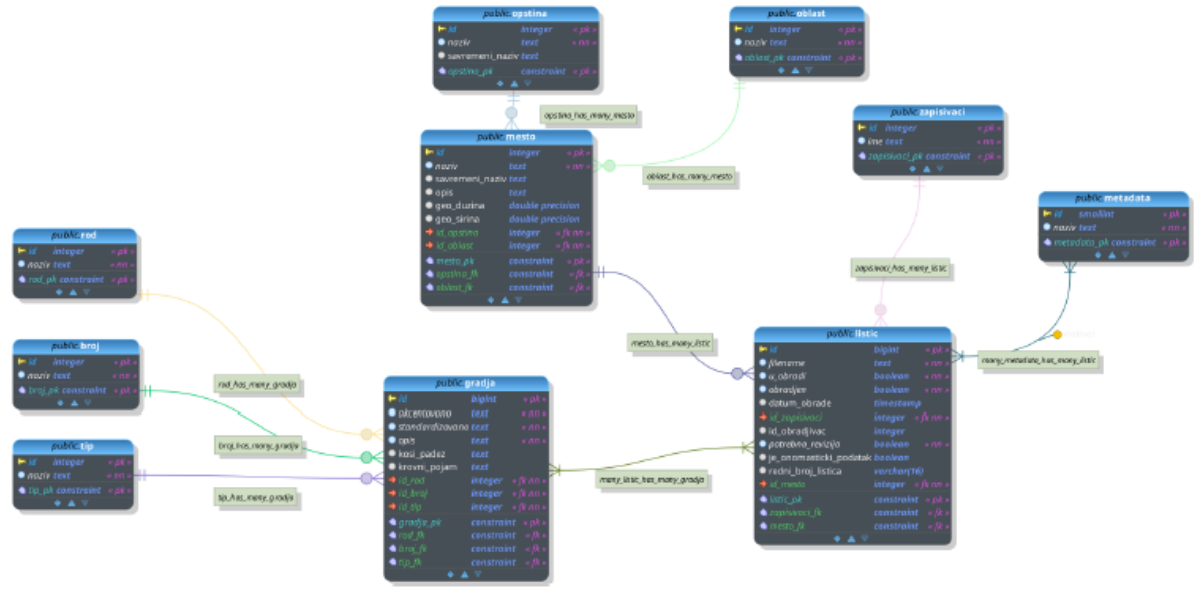
Дигитализација грађе одбора за ономастику САНУ – значај, циљеви и први кораци

У овом раду ће бити описан значај, циљеви, као и први кораци у процесу дигитализације ономастичке грађе Одбора за ономастику САНУ, која броји око 750 000 ономастичких података сакупљених са целокупног српског језичког простора у периоду од 1975. године до данас.

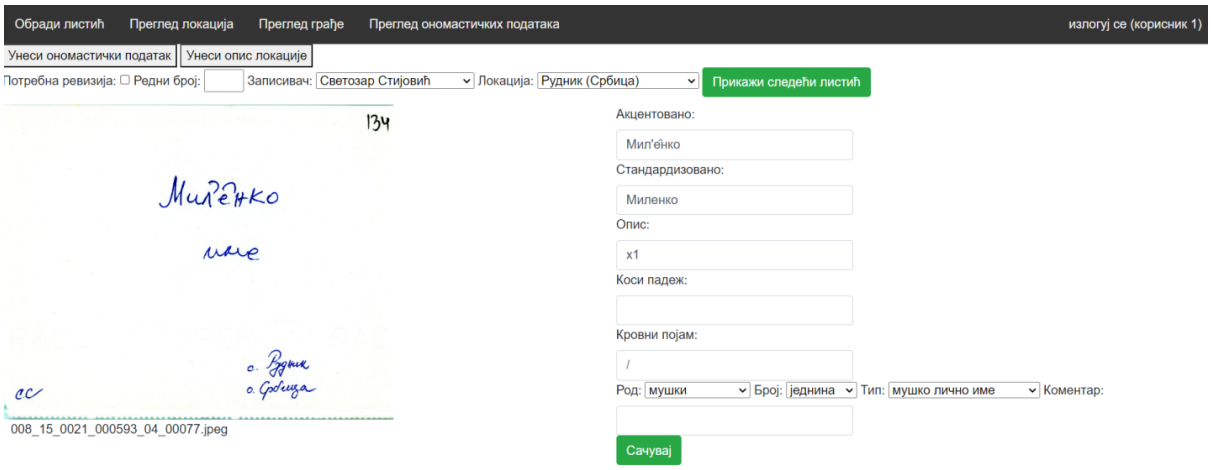
Опис првих корака у дигитализацији те грађе подразумева представљање припреме и процеса скенирања и документовања, затим моделирања различитих података записаних на листићима ономастичке грађе за потребе израде структуриране базе података (Слика 1) и креирања корисничког интерфејса за приступ бази и уношење података (Слика 2). Биће приказан поступак израде и имплементације интерфејса, уз осврт на проблеме и недоумице који су се јавили у процесу израде, као и на примењена решења.

Крајњи циљ дигитализације ономастичких записа јесте објављивање грађе у форми вишеструко претраживе платформе отвореног типа, доступне свима, како истраживачима тако и широј јавности. Платформа ће омогућити да се ономастички подаци организују, повежу и прикажу на више начина, укључујући различите врсте претрага и визуелизацију података у виду дигиталних мапа.

Кључне речи: *српски језик, дигитализација, ономастика, израда интерфејса, визуелизација података.*



Слика 1. Схема модела ономастичке грађе



Слика 2. Интерфејс за унос ономастичког податка

др Василије Милновић, др Александра Тртовац
Универзитетска библиотека „Светозар Марковић“, Београд
E-mail: {milnovic|aleksandra}@unilib.rs

проф. др Цветана Крстев, проф. др Ранка Станковић, проф. др Душко Витас
Друштво за језичке ресурсе и технологије - JePTex
E-mail: {cvetana|ranka|dusko}@jerteh.rs

Унапређење машинског разумевања текста и проналажења информација у историјским новинама у Србији

Развој и примена напредних језичких модела и технологија могу значајно побољшати прецизност проналажења информација у тексту скенираних новинских страница и њихово повезивање са базама знања доступним на интернету. С обзиром на све већу пажњу посвећену вештачкој интелигенцији у дигиталним хуманистичким наукама и архивским истраживањима, постоји потреба за развојем повезаних архивских података (Archival Linked Data). Иако технике вештачке интелигенције у истраживању и практичним решењима имају још много тога да понуде, оне пружају велики потенцијал дигиталним архивама за повећање обима повезаних података и иновирање начина приступа тим подацима. Аутоматизација процеса и смањење ресурса потребних за производњу повезаних архивских података омогућавају ефикаснији рад експерата, док је кључно да свака имплементација буде усмерена не само на њену употребу, већ и на њена ограничења, пристрасности и етичке импликације.

Колекција *Историјске новине* Универзитетске библиотеке „Светозар Марковић“ представља огроман и значајан ресурс који је био основа за неколико истраживачких и библиотечко-архивских пројеката. Да би се модернизовао приступ овим колекцијама текстова, ново истраживање се ослонило на активности ранијих, успешно спроведених, пројеката везаних за удаљено читање (*Distant Reading*). У овом раду биће представљено неколико иновација, односно сегмената унапређења семантичке видљивости историјских новина::

1. Унапређење претраге проширивањем упита тако да обухвати претрагу граматичким облицима кључне речи, ослањајући се на веб сервисе Друштва за језичке ресурсе и технологије и електронске речнике српског језика.
2. Проналажење, обележавање и екстракција кључних информација, као што су наслови, датуми, особе, организације и локације.
3. Идентификација важних ентитета и релација међу њима, користећи семантичке мреже као што су WordNet, Wikidata, GeoNames, као и векторске репрезентације речи развијене у оквиру пројекта ТЕСЛА (Text Embeddings - Serbian Language Applications).
4. Коришћење информација о локацијама за приказивање на карти локација поменутих у новинским чланцима, повезивање географских места са одређеним догађајима или временским одредницама.
5. Развој интерактивних визуализација које омогућавају корисницима да истражују и разумеју историјске информације из новинских чланака, користећи графове, карте и друге визуелне елементе.

Резултати који ће бити представљени треба да омогуће побољшано претраживање текста и екстракцију информација са скенираних новинских страница, визуализације које омогућавају истраживање историјских тема и догађаја, ефикасну претрагу и препоруке

корисницима на основу анализе садржаја новина, као и едукацију библиотекара и истраживача путем припремљених студија случаја.

Кључне речи: *језички модели, српски језик, дигитална хуманистика, архивски повезани подаци, вештачка интелигенција, удаљено читање, графови знања*

Андрија Сагић
Библиотека "Милутин Бојић", Београд, Начелник Одељења за дигитални развој
E-mail: andrija.sagic@milutinbojic.org.rs

Установе културе у ери вештаче интелигенције

Установе културе у својим фондовима баштине богат извор материјала (текстуална и фото грађа) који се могу користити у сврху машинског учења. Српски језик је класификован као језик са малим ресурсима. Највише доступних садржаја је доступно са веб садржаја који није у потпуности релевантан за тренирање модела јер садржи често лажне информације, (нпр. Милутин Бојић - песник је чувени фудбалер), такође већина информација на информативним порталима је "контаминирана" пропагандним садржајем и шокантним насловима који привлаче читаоце. Иницијатива "Цоллецтионс ас дата", покренута 2016. године на време је увидела значај вредности колекција које се налазе у установама културе. Дигитализација и припрема колекција за развој модела вештачке интелигенције, у данас може да постане важан сегмент у раду институција културе. Како можемо најбоље искористити дигиталне садржаје, како треба припремити колекције, које лиценце употребити приликом објављивања колекција и скупова података? Ово су питања на које је потребно дати одговоре у комуникацији са заједницом која развија моделе машинског учења.

На који начин могу установе културе да допринесу бољем развоју моделима вештачке интелигенције на језицима са мало доступним ресурсима, нпр. за генерисање текста, слика, звука, за аутоматску транскрипцију говора и других задатака?

Одговори на ова питања биће приказана у презентацији која ће упоредити светске праксе са домаћим подухватима.

Установе културе у свету годинама имају и дигиталне лабораторије које припремају и објављују скупове података из својих колекција, публикују алате за њихову припрему и даље коришћење, организују догађаје...

Да ли ће формирање прве дигиталне лабораторије у установи културе, следећи праксу колега из иностранства, моћи да реши, барем делимично, недостатак ресурса за тренирање модела? Надамо се да може.

Кључне речи: *дигитализација, културно наслеђе, машинско учење, установе културе*

Срђан Шуњур, Јелена Марковић
Универзитет у Источном Сарајеву, Филозофски факултет Пале,
Катедра за англистику
E-mail: {srdjan.sucur|jelena.markovic}@ffuis.edu.ba

Дигитализација српског књижевног наслеђа ијекавског изговора (1840–1920) при Центру за дигиталну хуманистику Филозофског факултета пале (прва фаза)

„Оно што није записано, и не постоји […]“, ријечи су којима Меша Селимовић започиње своју „Тврђаву“. У дигиталном окружењу, једна од интерпретација ових ријечи могла би бити: „Оно што није записано дигитално, и не постоји“. Значај и улога дигиталне хуманистике, између осталог, огледају у се конверзији аналогног у дигитално, тј. у дигиталном реформатирању. То и јесте један од задатака пројекта *Дигитализација српског књижевног наслеђа ијекавског изговора (1840-1920) при Центру за дигиталну хуманистику Филозофског факултета Пале (прва фаза)*.

Корпус репрезентативне књижевности писане на српском језику ијекавског изговора, чије је успостављање један од циљева пројекта, представља проширење корпуса SrpELTeC+ (из објективних разлога доминантно екавског изговора), чији је развој започет у оквиру завршене COST акције назива Distant Reading for European Literary History (COST Action CA16204). У тренутно доступним корпусима српског језика, ијекавски изговор није ни приближно представљен пропорционално својој заступљености код говорника српског језика као матерњег, те самим тиме ни доступан за корпусна истраживања, као ни ширу читалачку публику. Иницијално, замишљено је да корпус ијекавског изговора садржи 30 књижевних остварења, а одредница „прва фаза“ у називу пројекта наговјештава да су планирана и будућа проширења и надоградња корпуса.

У овом приказу представићемо пресјек реализације и проблема у реализацији 5 планираних фаза, и то: одабир текстова за дигитализацију, реализовање радионица у организацији Друштва за језичке ресурсе и технологије JePTeX, процес дигитализације одабраних текстова, публиковање припремљених дигитализованих облика на сајту Центра за дигиталну хуманистику при Филозофском факултету Пале, а у организацији Друштва за језичке ресурсе и технологије из Београда, као и бета тестирање.

Прва фаза, између осталог, обухвата приповијетке: Петра Кочића, Светозара Ћоровића, Васе Кондића, Марка Поповића, Милана Трифуновића, Јоаникија Памучине, потом путописе: Константина Хаџиристића, Јове Бесаровића, Саве Косановића, Ристе Бесаровића, Саве Пјешчића и Марка Цара, те романе: Светозара Ћоровића и Радована Тунгуз-Перовића.

За све дигитализоване публикације припремљени су и релевантни метаподаци, будући да је њихово постојање једна од предуслова за употребљивост корпуса у научне сврхе.

Кључне речи: *SrpELTeC+, дигитализација, српско књижевно наслијеђе, ијекавски изговор*

JуДиг радионице

Оља Перишић

Универзитет у Торину – Департман за стране језике и књижевност
https://unito.webex.com/meet/olja.perisic

Оља Перишић је доцент за српски језик при одсеку за Стране језике и књижевност Универзитета у Торину. Стекла је докторат из области дигиталне хуманистике 2020. године на Универзитету у Ђенови. Има вишегодишње искуство као књижевни преводилац и тумач за италијански и српски језик. Основна поља истраживања су јој примена корпуса и дигиталних технологија у настави српског као страног језика, превођење, билингвална лексикографија и контрастивна граматика. Ауторка је монографије Il corpus per imparare il serbo. Il futuro dell'apprendimento linguistico (2023).

Претрага корпуса (CQL): Lexical gaps у двојезичним корпусима

Радионица је намењена предавачима страних језика, истраживачима, преводиоцима и свима онима који су заинтересовани за коришћење језичких корпуса у настави страних језика и превођењу. Претходно искуство у области корпусне лингвистике није неопходно. Учесници ће се упознати са методолошким основама корпусне лингвистике и начинима њихове примене у различитим областима истраживања.

Посебно ћемо се осврнути на међујезичне лексичке неуједначености које се тичу непостојећих еквивалената у једном од језика обично условљених културолошким разликама и језичким анизоморфизмом (полисемија и lexical gaps).

Циљ радионице је да преко практичних вежби покаже учесницима како да критички претражују паралелне корпусе почев од основних, преко сложених упита који обухватају и обележене ентитете (топоними, антропоними итд.). У другом делу учесници ће се опробати у самосталном претраживању корпуса, екстракцији преводних еквивалената и ентитета у двојезичним корпусима књижевних текстова It-Sr-NER и SerbItaCor3_sr.

Тим пројекта ТЕСЛА: Милица Иконић Нешић, Михаило Шкорић и Саша Палинкар

Универзитет у Београду – Филолошки факултет, Универзитет у Београду – Рударско-геолошки факултет и Друштво за језичке ресурсе и технологије JePTex

Препознавање именованих ентитета и повезивање са Википодацима

Радионица ће пружити учесницима увид у технике аутоматског препознавања именованих ентитета (Named Entity Recognition - NER). Полазници ће научити како се примењују модели којима се идентификују личности, места и организације у књижевним делима и повезују са одговарајућим ентитетима на Википодацима.

Практични део радионице ће укључити коришћење алата и модела, укључујући оне који се заснивају на векторској репрезентацији речи а који настају у склопу пројекта ТЕСЛА Векторизација текста – апликације за српски језик (Text Embeddings – Serbian Language Applications, PRIZMA #7276) који финансира Фонд за науку Републике Србије. Користиће се алати и сервиси доступни на https://ners.jerteh.rs/ као и модел jertex-355-тесла, алат INCEPTION и Википодаци. Полазници ће се упознати и са осталим ресурсима који се развијају у склопу пројекта ТЕСЛА (https://tesla.rgf.bg.ac.rs/).

Тим пројекта ТЕСЛА: Ранка Станковић, Цветана Крстев и Душко Витас
Универзитет у Београду – Рударско-геолошки факултет и Друштво за језичке ресурсе и технологије JePTex

Анализа корпуса: текстометрија, TXM и други алати

Радионица је намењена свима које интересују савремене технике и методе у обради природних језика. Полазници ће се прво упознати са концептом и методама текстометријске анализе уграђеним у алат TXM, а потом и моделима и ресурсима које је за српски језик развило Друштво за језичке ресурсе и технологије JePTex.

Циљ радионице је да покаже полазницима како могу да користе текстометријску анализу над готовим корпусима Јертех-а, а потом и да креирају сопствене корпусе. Други део радионице биће посвећен креирању и текстометријској анализи сопствених корпуса коришћењем алата TXM. За вежбе ће бити припремљени текстови из корпуса српских романа (1840–1920) SrpELTeC и паралелни корпус резимеа са конференције ЈуДиг. Полазници ће се упознати са ресурсима који се развијају у склопу пројекта ТЕСЛА (Text Embeddings – Serbian Language Applications, PRIZMA #7276), који финансира Фонд за науку Републике Србије (https://tesla.rgf.bg.ac.rs/).

Benedikt Perak, docent

Faculty of Humanities and Social Sciences, University of Rijeka

Dragana Špica, docent

Cultural Studies Department University in Pula, Croatia
https://portal.uniri.hr/Portfelj/1078

Benedikt Perak je docent na Filozofskom fakultetu Sveučilišta u Rijeci, gdje predaje kolegije poput Data Science u kulturi, Alati i metode digitalne lingvistike te Umjetna inteligencija i komunikacija. Također je voditelj mikrokvalifikacije Jezične tehnologije: Analiza teksta i ekstrakcija informacija. Njegov rad uključuje primjenu velikih jezičnih modela (LLM) u analizi jezičnih mreža, kao i korištenje naprednih digitalnih alata za obradu prirodnog jezika i računalnu semantiku. Perak ima bogato istraživačko iskustvo u području digitalne lingvistike, osobito u primjeni modela za analizu leksičkih mreža i računalnu obradu jezika. Aktivno koristi računalne programe poput Pythona, Networkx i Neo4j za obradu i vizualizaciju jezičnih podataka, stvaranje jezičnih korpusa, te razvoj alata za računalnu lingvistiku.

Dragana Špica je docent na Odsjeku za azijske studije Filozofskog fakulteta Sveučilišta Jurja Dobrile u Puli gdje predaje japanski jezik i povezane predmete na prijediplomskom studiju Japanski jezika i kultura i diplomskom studiju Japanologija. Autorica je više znanstvenih radova iz oblasti lingvistike te suautorica monografije Uvod u znanost o japanskom jeziku, prve takve vrste na području bivše Jugoslavije. Zanima je japanska fonologija, pridjevi u japanskom te leksikologija. Diplomirala je u Beogradu, a magistrirala i doktorirala u Osaki.

Korištenje velikih jezičnih modela za stvaranje leksičkih mreža s fokusom na ekstrakciju sinonima

Ova radionica ima za cilj upoznati sudionike s tehnikama korištenja velikih jezičnih modela (GPT-4) za automatiziranu ekstrakciju sinonima i antonima te građenje leksičkih mreža. U sklopu radionice, sudionici će naučiti kako pravilno postaviti promptove za jezične modele, definirati leksičke odnose, te koristiti rezultate za vizualizaciju i analizu semantičkih struktura.

Glavne teme radionice:

- Uvod u koncepte leksičkih mreža i relacija (sinonimija, antonimija, hijerarhijski odnosi)
- Korištenje LLM-ova za ekstrakciju leksičkih relacija
- Praktične demonstracije prompt-engineeringa za dobivanje preciznih leksičkih podataka
- Analiza rezultata i vizualizacija leksičkih mreža pomoću graf modela
- Praktične primjene u NLP-u i leksikografiji

Ciljevi:

- Povećanje točnosti i brzine ekstrakcije leksičkih relacija kroz korištenje GPT-a
- Razvijanje vještina prompt-engineeringa za potrebe leksikografskih istraživanja
- Kreiranje i evaluacija semantičkih grafova na temelju dobivenih podataka

Radionica je namijenjena istraživačima i praktičarima koji se bave lingvističkim analizama, leksikografijom, te prirodnim jezičnim procesiranjem (NLP), no nije potrebno duboko tehničko predznanje.

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

811.163'322(048)(0.034.2)
004.8(048)(0.034.2)

МЕЂУНАРОДНА конференција Јужнословенски језици у дигиталном окружењу ЈуДиг (2024 ; Београд)

Зборник резимеа [Електронски извор] / Међународна конференција Јужнословенски језици у дигиталном окружењу ЈуДиг, 21-23. новембар 2024, [Београд] ; [организатори Филолошки факултет Универзитета у Београду [и], Друштво за језичке ресурсе и технологије (JePTex)]. - Београд: Филолошки факултет Универзитета, 2024 (Београд : Универзитет, Филолошки факултет). - 1 електронски оптички диск (CD-ROM) : текст ; 12 cm

Системски захтеви: Нису наведени. - Насл. са насловног екрана. - Текст ћир. и лат. - Тираж 100.

ISBN 978-86-6153-755-4

а) Јужнословенски језици -- Рачунарска лингвистика – Апстракти

COBISS.SR-ID 157061385