UDK 811.163.41’344:004.853
811.161.1’344:004.855.6

Иновације у настави, XXXVII, 2024/2, стр. 156–167
DOI: 10.5937/inovacije2402154M

Рад примљен: 23. 2. 2024.
Рад прихваћен: 22. 5. 2024.

Кратки научни прилог

Лука М. Меденица[1]
Универзитет у Београду, Филолошки факултет, Београд, Србија

Милена С. Опарница
Истраживачко-развојни институт за вештачку интелигенцију, Нови Сад, Србија

Креирање алгоритма за откривање лексичких минимума у српском као страном језику на нивоу А1

Резиме: *Програми за аутоматску корекцију и адаптацију текста имају све већи утицај на савремену лингводидактику. Имајући у виду да се такви алати не користе у настави српског као страног језика, циљ нашег рада био је одређивање лексичких минимума за ниво А1 у српском као страном језику и креирање алгоритма за њихово откривање, односно предузимање првих корака за развој програма за аутоматску корекцију и адаптацију текста. У истраживању смо користили методологију која се ослања на лингводидактику руског као страног језика, имајући у виду развијене корпусе лексичких минимума у том словенском језику и постојање онлајн-алата Текстометр за проверу сложености текста. Тако смо дошли до 783 лексеме, које представљају списак лексичких минимума за ниво А1 у српском као страном језику. Затим смо за потребе рада креирали алгоритам у програмском језику Python, који смо испитали на конкретном тексту и установили одређене недостатке када је лематизација текста у питању. У наредном периоду је потребно извршити дораду лематизатора ради развоја програма за аутоматску корекцију текста.*

Кључне речи: *српски као страни језик, ниво А1, лексички минимуми, аутоматска корекција и адаптација текста*

1 luka.medenica@fil.bg.ac.rs;
https://orcid.org/0000-0002-4453-0703

Увод

Убрзани технолошки напредак и развој вештачке интелигенције имају све снажнији утицај на наставу и учење страних језика (Alharbi, 2023). Сведоци смо чињенице да се у појединим великим језицима, као што су енглески, шпански и руски, улажу огромна финансијска средства у развој лингводидактике, па тако и у програме за аутоматску корекцију и адаптацију текста (Alarcon et al., 2019; Burstein et al., 2007; Laposhina & Lebedeva, 2021). Један од циљева ових програма јесте брзо и ефикасно поједностављивање текста за потребе наставе и учења конкретног страног језика. Оваква адаптација од велике је користи, како наставницима, који су у прилици да дидактизују текст за потребе конкретне групе полазника, тако и особама које усвајају језик, јер оне тако имају могућност да поједноставе доступне текстове врло лако и поуздано. Међутим, аутоматска адаптација је доста сложен процес који захтева синергију неколико важних компоненти. Тако поступку измене и прилагођавања текста претходи неколико значајних корака које је потребно предузети ради креирања таквог програма.

Анализа постојећих дигиталних алата

На самом почетку неопходно је упознати се са могућностима таквих алата, те ћемо у наредним пасусима детаљно размотрити један од активних програма који би могао послужити за развој и даљи напредак лингводидактике српског као страног језика.

Текстометар (Laposhina & Lebedeva, 2021) представља пример дигиталног алата који служи за прецизно одређивање нивоа сложености текста за руски као страни језик, а према Заједничком европском референтном оквиру за живе језике (Council of Europe, 2020). Имајући у виду да је у питању словенски језик који има много више сличности са српским него енглески или шпански, што због споре промене базичне лексике, коју чини прасловенска лексика, што због утицаја рускословенског и руског на славеносрпски и савремени српски језик (Dragićević, 2018), *Текстометар* се чини као веома добар лингводидактички путоказ за прве кораке у изради програма за аутоматску корекцију и адаптацију текста када је српски као страни језик у питању.

У тренутној фази развоја овај алат није у стању да понуди аутоматску адаптацију текста, већ даје списак свих употребљених лексема, фреквентност њихове употребе у тексту и распоређује их према језичким нивоима (од А1 до Ц2), у складу са лексичким минимумима и корпусом, који је састављен од великог броја уџбеника за руски језик као страни. Тако алат пружа информацију о нивоу сложености текста према скали Заједничког европског референтног оквира за живе језике (Council of Europe, 2020), при чему корисник добија информацију о проценту покривености текста лексиком конкретног језичког нивоа, али и низ других веома корисних информација, попут броја јединствених речи, описне сложености текста у оквиру сваког од нивоа и сл. (Laposhina & Lebedeva, 2021). На основу свих наведених података наставник страних језика у прилици је да адаптира текст и затим га поново унесе у програм како би проверио његов ниво сложености, односно прилагођава га потребама својих полазника.

Методолошки оквир истраживања

На основу детаљне анализе дигиталног алата *Текстометар* закључили смо да је за потребе креирања програма за аутоматску корекцију и адаптацију текста на српском као страном језику најпре потребно размотрити питање лексема које ће ући у корпус за сваки од нивоа према Заједничком европском референтном оквиру за живе језике (Council of Europe, 2020). Међутим,

питање лексичких минимума за сваки од нивоа у великој мери превазилази оквире овога рада, нарочито ако се узме у обзир да је ова област у српском као страном језику недовољно истражена и веома сложена. Узимајући ову чињеницу у обзир, као и комплексност програма за аутоматску корекцију и адаптацију текста, циљеви овог истраживачког рада су: 1) одређивање лексичких минимума за ниво А1 у српском као страном језику, односно стварање корпуса за потребе рада програма; 2) креирање алгоритма за откривање лексичког минимума, а за потребе текста на српском као страном језику, односно употребу таквог програма на конкретном тексту.

Одређивање лексичких минимума за ниво А1

Први корак у креирању корпуса за потребе програма за корекцију и адаптацију текста представља одређивање лексичких минимума за сваки од нивоа, почев од најнижег. На основу таквог корпуса могао би се установити ниво сложености текста који корисник унесе у програм. Осим тога, постојала би могућност за одређивање оних лексема које не припадају задатом нивоу, те би се олакшали наредни програмски кораци ради аутоматске корекције и адаптације текста. Српски као страни језик у овом тренутку нема званичне спискове лексичких минимума, а на основу увида у стручну и научну литературу закључујемо да овом питању није посвећено превише пажње.

У једном недавном истраживању описан је покушај креирања дела списка лексичких минимума за ниво А1, који је резултирао корпусом од 325 лексема (Medenica, 2023). Будући да је ова методологија дала конкретне резултате у виду корпуса који је неопходан за креирање програма за аутоматску корекцију и адаптацију текста, одлучили смо да поновимо наведено истраживање, с тим да овога пута обухватимо већи број анализираних речи, а ради креирања потпуног списка лексичких минимума за ниво А1.

У првом кораку коаутор рада је саставио нови текст степеноване лектире за потребе наставе и учења руског као страног језика на Филолошком факултету Универзитета у Београду[2], на основу чега смо анализирали нових 6309 речи у руском језику (у претходном истраживању је анализирано 3047 речи). Приликом састављања текста степеноване лектире аутор се руководио важећим списком лексичких минимума за руски као страни језик, који за ниво А1 броји 780 лексема (Andriushina & Kozlova, 2014). Затим је текст унет у дигитални алат *Текстометр*, који је одредио да је корпус нивоа А1, односно да 96% лексема припада нивоу А1 према Заједничком европском референтном оквиру за живе језике (Council of Europe, 2020). Тако су и ручна и дигитална провера текста указале да је у питању корпус који према сложености припада првом језичком нивоу, уз минимална одступања од 4%.

У другом кораку приступили смо превођењу наведене степеноване лектире на српски језик, прилагодивши је духу језика. Као резултат настао је текст од 6805 речи (очигледна разлика између руског и српског језика од скоро 500 речи пре свега се објашњава употребом енклитика у српском језику). Наведени корпус смо детаљно анализирали у програму AntConc (верзија 4.1.1) и објединили га са 325 лексема које смо добили у претходном истраживању како не би дошло до понављања. Осим тога, изоставили смо из корпуса 4% речи које је *Текстометр* одредио као речи које не припадају нивоу А1, а које због потребе кохерентне и кохезивне приче није могуће изоставити из самог текста степеноване лектире (*аґент*, *ванила, водич, романтичан* и сл.). Добијени корпус смо поредили са још два ради његове додатне корекције. Први корпус представља предлог списка лексичких минимума за српски као страни језик и броји 1920 лексема (Krajišnik, 2016). Овај списак Весне Крајишник

2 Степенованој лектири на руском („Ивановы – часть вторая" и „Ивановы – часть третья") анализираној у раду могуће је приступити путем линка https://storiesbyluka.com/.

представља једини предлог лексичких минимума који је познат стручној јавности, али који по броју лексема знатно премашује оквире нивоа А1. Други корпус је актуелни списак лексичких минимума за ниво А1 у руском као страном језику, и он броји 780 лексема (Andriushina & Kozlova, 2014). Тако смо додатно кориговали корпус, додавши поједине лексеме које нису биле употребљене у самој степенованој лектири, а које су недостајале у погледу употпуњавања целина (недостајали су поједини дани у недељи, неке основне боје, поједини бројеви и називи месеци, националности за мушки или женски род и сл.). На крају смо добили списак од следеће 783 лексеме, што је бројчано готово идентично списку лексичких минимума у руском језику од 780 лексема за ниво А1.

Табела 1. Списак лексичких минимума за српски језик као страни, ниво А1.

а	Божић	вода	дати
август	боја	воз	два
авион	болестан	возити	двадесет
адреса	болети	волети	дванаест
аеродром	болница	воће	деведесет
Александар	Бразил	врата	девет
Александра	браон	вратити се	деветнаест
али	брат	време	девојка
Ана	брзо	где	девојчица
апликација	број	географија	деда
апотека	булевар	гитара	део
април	важно	глава	десет
аутобус	Ваш	главни град	десно
аутомобил/ауто	вежбати	гладан	дете
аутор	велик	гласно	децембар
баба	веома	гледати	дечак
бавити се	веровати	глумац	динар
базен	вест	глумица	дневна соба
бака	већ	глуп	до
балет	вече	говорити	добар
банана	вечера	година	добити
банка	вечерас	гост	добро
био	вечерати	град	добродошао
Београд	ви	грешка	довиђења
библиотека	Ви	група	док
бизнисмен	видети	да	доктор
биоскоп	виљушка	да	документ
бити	висок	далеко	дом здравља
бицикл	власник	дан	домаћи
близу	власница	данас	допадати се

доручак
доручковати
досадан
доћи
драго
дрво
држава
држати
друг
другарица
други
друштвена мрежа
дуго
ево
евро
Европа
електронски
енглески
ето
жао
жедан
желети
жена
живети
живот
жут
за
заборавити
заборављати
задатак
заједно
закаснити
замолити
занимљив
занимљиво
запамтити
зар
затворити
затим
зато
заузет
зашто
због
звати
звати се
зграда
здрав
здраво
зелен
зид
зима
Златибор
знати
и
Иван
Ивана
игра
играти
из
иза
изаћи
извинити
или
имати
име
имејл
Индија
инжењер
институт
интернет
ипак
испит
испред
испричати
исти
истина
историја
Италија
ићи
ја
јабука
јавити
јаје
јак
јакна
јануар
Јапан
је
један
једанаест
једини
једном
језик
Јелена
јело
јер
јесам
јесен
јести
јефтино
још
јул
јун
јутро
јуче
к(а)
када
казати
какав
како
календар
као
капа
карта
касније
каснити
касно
кауч
кафа
кафић
кашика
килограм
километар
Кина
Кинез
кинески
Кинескиња
киша
кључ
књига
књижара
књижевност
ко
код
који
колико
компјутер
композитор
комшија
комшиница
Копаоник
кошарка
коштати
кошуља
кревет
кренути
кроз
кромпир
кћерка/ћерка
кувати
куда
купатило
купити
куповати
кућа
кухиња
лак
лако
лево
лежати
лек
лекар
леп
лепо
лето
лифт
лице
Лондон

лош
лоше
Лука
мај
мајица
мајка
мали
мало
мама
Марко
март
математика
мачка
менаџер
месец
месо
метро
ми
Милица
минут
мислити
млад
младић
много
мобилни
можда
мој
молити
морати
море
Москва
моћи
муж
музеј
музика
музичар
мушкарац
на
надати се
назад
налазити се
наочаре
написати
напоље
напољу
направити
наравно
наставник
наставница
наћи
наука
наш
не
недеља
неки
неко
неколико
немати
Немац
Немачка
немачки
Немица
нешто
ни
нигде
низак
није
никада
нико
нисам
Ниш
ништа
нов
новац
новембар
Нови Сад
новине
нога
нож
нормално
нос
носити
ноћ
нула
његов
њен
њихов
о
обавезно
обично
обућа
обући
овај
овако
овде
од
одавно
одакле
одбојка
одговарати
одговор
одговорити
одећа
одједном
одличан
одлично
одлучити
одмарати се
одмах
одморити се
око
око
октобар
омиљен
он
она
онај
онда
оне
они
оно
опет
ормар
осам
осамдесет
осамнаест
осећати се
осми
основна школа
отац
отворити
отићи
падати
пажљиво
паметан
панталоне
Париз
парк
паркинг
пас
пасош
певати
певач
певачица
педесет
пекара
песма
пет
петак
пети
петнаест
пешке/пешице
пиво
пијаца
писати
писац
писмо
питање
питати
пити
плав
план
планина
платити
по
поврће
погледати
позвати

поздравити
познавати
познат
позориште
показати
поклањати
поклон
поклонити
полако
полица
помагати
помислити
помоћ
помоћи
понедељак
понекад
поред
породица
порука
посао
послати
после
последњи
поставити
постати
потребан
почети
почињати
пошта
право
празник
прати
први
прво
пре
превести
преводити
предавање
председник
предсобље
презивати се
презиме
преселити се
приземље
пријатељ
пријатељица
пријатно
пример
прича
причати
проблем
продавница
прозор
пролеће
проћи
професор
професорка
прочитати
прошли
психологија
пуно
пуњач
пут
путовати
радио
радити
радник
радо
разговарати
размислити
размишљати
разред
разумети
ранац
рано
ред
резултат
река
ресторан
ретко
рећи
реч
риба
родитељ
родити се
рођендан
розе
рука
Рус
Русија
руски
Рускиња
ручак
ручати
с(а)
сав
сад(а)
салата
сам
само
сарма
сат
сваки
све
свеска
свет
свидети се
свиђати се
свирати
свој
седам
седамдесет
седамнаест
седети
село
септембар
серија
сести
сестра
сетити се
сећати се
сигурно
син
синоћ
сир
сјајан
сјајно
скоро
скупо
слава
славити
сладолед
слати
следећи
слика
слободан
слово
слушати
смејати се
снег
соба
сок
спавати
спаваћа соба
споменик
спор
споро
спорт
спортиста
спрат
спремати
Србија
Србин
среда
срести
сретати
срећа
срећан
Српкиња
српски
стадион
стајати
стално
стан
станица
стар
стварно
Стефан

сто
сто
столица
стомак
страна
странац
стрина
стриц
студент
студенткиња
студирати
субота
сувенир
сунце
сутра
схватити
тад(а)
тај
тако
такође
такси
тамо
тањир
тата
тачан
тачно
ташна
твој
тежак
телевизија
телевизор
телефон
температура
тенис
тераса
тетка
теча
тешко
ти
тихо
тоалет
толико
топло
торба
тражити
трамвај
требати
тренерка
трећи
три
тридесет
тринаест
тролејбус
трпезарија
трчати
ту
туриста
туристички
туристкиња
туширати се
ћао
у
у реду
увек
увече
узети
узимати
ујак
ујна
ујутро
укусан
улица
уместо
умети
уморан
умрети
универзитет
унук
унука
уопште
уписати
упознати
урадити
Ускрс
устајати
устати
утакмица
уторак
ученик
ученица
учитељ
учитељица
учити
уџбеник
факултет
фебруар
физика
филм
фотеља
фотографија
фотографисати
Француз
француски
Францускиња
фудбал
фудбалер
хајде
хало
хаљина
хвала
хиљада
хитна помоћ
хладан
хладно
хлеб
хоби
ходник
хотел
храна
хтети
цвет
цена
центар
цео
ципеле
црвен
црн
цртати
чај
чарапа
час
чаша
чекати
честитати
често
четвртак
четири
четрдесет
четрнаест
чији
читати
човек
чоколада
чути
џемпер
шалити се
шампион
шездесет
шеснаест
шест
шетати
шећер
школа
Шпанац
Шпанија
шпански
Шпањолка
шта
што

Иако је овај списак лексичких минимума од 783 лексеме тренутно једини доступан списак за српски као страни језик за ниво А1 према Заједничком европском референтном оквиру за живе језике (Council of Europe, 2020), он није званичан и треба га узети с дозом опреза (нарочито у погледу одабира конкретних назива националности и властитих имена, али и других важних питања). Ограничења понуђеног лексичког минимума се пре свега огледају у методологији, која се ослања на лингводидактику руског језика као најближег језика са развијеним корпусима лексичких минимума за страни језик. Убудуће су потребна додатна истраживања која ће бити посвећена детаљној анализи свих лексема овог корпуса ради евентуалне корекције и допуне списка, односно креирања коначне званичне верзије списка лексичких минимума за ниво А1. На основу њега било би могуће прећи на више језичке нивое, а шта би значајно унапредило лингводидактику српског језика као страног, а самим тим убрзало и развој дигиталних алата, а шта у овом тренутку превазилази оквире нашег рада.

Ипак, упркос свим могућим недостацима и ограничењима, верзија лексичких минимума од 783 лексеме у српском као страном језику за ниво А1, а до које смо дошли у раду, у потпуности задовољава потребе нашег истраживања, односно даје нам могућност за тестирање алгоритма за рад програма за аутоматску корекцију и адаптацију текста, тј. његове провере на конкретном тексту.

Тренутне могућности програма за аутоматску корекцију и адаптацију текста

Као први корак при развијању програма који би имао могућности за аутоматску корекцију и адаптацију текста на српском језику креирали смо једноставни алат у програмском језику *Python,* чији је задатак да обележи речи које не припадају нивоу А1, у складу са речником дефинисаним у овом раду. Користили смо *classla* библиотеку за аутоматску анотацију и лематизацију текста (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023). Свакако, с обзиром на то да смо до сада говорили само о лексемама, треба напоменути да се у корпусној и рачунарској лингвистици за основни облик речи користи термин *лема*. Лематизација подразумева свођење текста на појединачне леме, тј. основне облике појединачних речи. На пример, једна лексема *дневна соба* биће лематизована као две леме: *дневни* и *соба*. Након изведене аутоматске лематизације, програм проверава да ли се леме у тексту налазе у наведеном речнику нивоа А1. Исход овог програма представља листу лема које се не налазе у речнику А1, као и текст у којем су обојене леме које не припадају овом речнику.

За тестирање овог алата одабрали смо причу *Петровићи*[3]. Ова прича укупно садржи 4109 токена (у питању су засебне речи и интерпункцијски знаци) и 463 различите леме. У наставку је приказан део текста који је коришћен за анализу и тестирање (обележене речи не припадају речнику):

Vidimo se . Čeka me drugarica . Dušan : Ćao ! Dušan i Marija su komšije odavno , ali Dušan već godinu dana sve vreme misli na Mariju . Da **li** je to **ljubav** ? On misli da **voli** Mariju . Marija ne zna da je Dušan **voli** . Ona misli da su samo komšije . Tatjana je Dušanova sestra . Svi je zovu **Tanja** . **Tanja** ima 14 (četrnaest) godina . **Tanja** je učenica 8 (osmog) razreda . **Tanja** želi da bude kao Marija . **Tanja** misli da je Maja veoma pametna i lepa . Tanji se dopadaju (= sviđaju) **Marijine** haljine , Marijine **tašne** , **Marijina** obuća . **Tanja** želi sve to da kupi . **Tanja** zna da Dušan **voli** Mariju . A Tanji se dopada Marko . Marko ima 15 (petnaest) godina . Marko je lep . Marko je veoma lep . On je veoma **talentovani** sportista , **prvak** Srbije u **boksu** . **Tanja voli** i svog psa

3 Степенованој лектири *Петровићи* могуће је приступити путем линка https://storiesbyluka.com/. Аутор наведене степеноване лектире је уједно и коаутор рада.

. Svaki dan ide sa njim u park . Pas je veoma pametan , zove se Snupi . Jutro je , 9 časova . **Tanja** ide u park sa svojim psom Snupijem . **Tanja voli** da se šeta . U parku ima puno pasa . A u parku je i Marko , jer i on ima psa . **Tanja** : Ćao ! Marko : Ćao , ja sam Marko . A kako se ti zoveš ? **Tanja** : Marko ! Ne znaš kako se zovem ? **Neko** te je **udario** u glavu ? Šta da radim ? Marko , sviđaš (= dopadaš) mi se . Naravno , **Tanja** je to pomislila . **Tanja** : Marko , kako se osećaš ? Ali Marko se smejao . I **Tanja** je shvatila … Marko zna kako se ona zove . Marko misli da je to smešno . **Tanja** se nije smejala , ali Marko se smejao . Marko se smejao sat vremena . Marko : **Ha-ha-ha** ! To je **baš smešno** ! To je **baš** , **baš smešno** ! Marko **voli** boks . Marko je šampion ! Marko je lep . Markov tata svaki dan govori Marku : – Mi smo šampioni ! Marko odgovara : – Ja sam šampion ! Markova sestra takođe govori Marku svaki dan : – Marko , bićemo **svetski** šampioni ! Marko odgovara : – Ja ću biti **svetski** šampion ! Markova mama govori Marku svaki dan : – Marko , moraš da učiš ! Marko ne **voli** da uči . Marko **voli** da bude šampion . Marko ne **voli** istoriju , matematiku , geografiju , strane jezike . Marko **voli** samo sport ! Marko **voli** da se šeta sa svojim psom.

Листа свих лема које алгоритам не препознаје као део корпуса лексичких минимума који смо користили за ниво А1:

['Dušan', 'voliti', 'danju', 'bravo', 'talentovan', 'mašta', 'psiholog', 'tko', 'ljubav', 'novina', 'kralj', 'šesti', 'otvarati', 'vanila', 'Dušanov', 'Tanja', 'Marijin', 'prvak', 'boks', 'netko', 'udariti', 'smešan', 'Ha-ha-ha', 'baš', 'Markov', 'svetski', 'stran', 'šala', 'nitko', 'romantičan', 'romantika', 'vrnjački', 'banja', 'cel', 'muzički', 'rešavati', 'misao', 'ljubavni', 'matematičar', 'Žanina', 'odlučivati', 'cveće', 'neromantičan', 'razmileti', 'potom', 'mamin', 'dnevni', 'sediti', 'mobilan', 'torta', 'tašan', 'mamiti', 'dopasti', 'dolaziti', 'stol', 'Goran', 'naredan', 'čokoladni', 'prav', 'rešiti', 'Žanin', 'davati', 'Tanjin', 'bel', 'Jeca', 'jecati', 'Ječin', 'otkud', 'mirno', 'puta', 'Puškin']

У оквиру листе лема које не припадају нивоу А1 нашле су се лексеме које се не налазе у речнику, као што су *таленшован*, *машша*, *романшичан* и сл. Такође, можемо приметити погрешно лематизоване речи, као што су „волети" (лематизовано као *волиши*), „ко" (лематизовано као *шко*), „неко" (лематизовано као *нешко*), „нико" (лематизовано као *ништко*), „ташна" (лематизовано као *ташан*), „Јеца" (лематизовано као *јецаши*) и др. Укупно, од 463 леме, 71 лема припада листи лема које се не налазе у речнику. Од тога, 17 лема је погрешно лематизовано, што чини 3,67% укупног броја лема. Ипак, важно је напоменути да ово не искључује остале погрешно лематизоване речи. Нисмо анализирали оне леме које су се нашле у речнику, те се могло десити да је нпр. именица *мисао* лематизована као глагол *мислиши*.

Исти текст (*Пешровићи*, 463 јединствене леме) лематизовали смо и помоћу алата у оквиру којег је имплементиран *TreeTagger* (Stanković et al., 2020; Stanković i Škorić, 2021). У листи лема које не припадају речнику овога пута се нашло 108 лема, док је 35 од укупног броја лема погрешно лематизовано (7,56%). Ипак, овде се ради о лематизатору који прво додељује највероватнију врсту речи, а затим додељује лему. Због тога, десило се то да је иста лексема *мама* лематизована као *мама, мам* и *мамиши*. Такође, важно је напоменути да је овај лематизатор треniран на ограниченом скупу корпуса који обухвата текстове из области права, здравства, образовања, као и поједине новеле из прве половине двадесетог века.

Закључак

Имајући у виду да у времену великог технолошког напретка за српски као страни језик још увек не постоје адекватни алати за аутоматску адаптацију текста, циљ овог рада био је да одредимо лексичке минимуме за ниво А1 у српском као страном језику и покушамо да креирамо алгоритам за откривање наведеног лексичког минимума.

Методологија која је коришћена за добијање списка лексичких минимума се понајвише ослања на лингводидактику руског као страног језика, који има развијене корпусе лексичких минимума, као и онлајн-алат *Текстометр*. Уз све потенцијалне недостатке наведене методологије и ограничења која смо описали у раду, дошли смо до списка од 783 лексеме за ниво А1 према Заједничком европском референтном оквиру за живе језике (Council of Europe, 2020). Овај корпус је у потпуности задовољио први циљ нашег истраживања, односно омогућио нам је тестирање алгоритма за рад програма за аутоматску корекцију и адаптацију текста. Ипак, у наредном периоду би било корисно додатно истражити наведени корпус, односно анализирати све лексеме ради евентуалне корекције и/или допуне списка. Осим тога, потребно је размотрити и креирање уџбеничког корпуса српског као страног језика за ниво А1, који би могао значајно унапредити списак лексичких минимума и рад програма за аутоматску корекцију и адаптацију текста.

Када је реч о другом циљу, алгоритму за откривање лексичког минимума, за потребе нашег истраживања креирали смо једноставни алат у програмском језику *Python*, затим смо у алат унели текст и покушали да откријемо лексеме које припадају лексичком минимуму за ниво А1. За аутоматску анотацију и лематизацију текста користили смо *classla* библиотеку и *TreeTagger* и установили одређене недоследности у њиховом раду када је лематизација у питању, односно забележили смо 3,67% и 7,56% погрешно лематизованих речи. Имајући све наведено у виду, за потребе креирања алгоритма за откривање лексичких минимума неопходно је извршити још неколико важних корака који би обухватили евалуацију и дораду лематизатора како би програм могао да даје тачне информације о сложености текста, а, коначно, и ради извршавања адекватне аутоматске корекције и адаптације текста.

Литература

- Alarcon, R., Moreno, L., Segura-Bedmar, I., & Martinez, P. (2019). Lexical simplification approach using easy-to-read resources. *Procesamiento de Lenguaje Natural*, *63*, 95–102. https://doi.org/10.26342/2019-63-10
- Alharbi, W. (2023). AI in the Foreign Language Classroom: A Pedagogical Overview of Automated Writing Assistance Tools. *Education Research International*, 1–15. https://doi.org/10.1155/2023/4253331
- Andriushina, N. P, & Kozlova, T. V. (2014). *Leksicheskii minimum po russkomu iazyku kak inostrannomu. Elementarnyi uroven. Obshchee vladenie*. Zlatoust.
- Burstein, J., Shore, J., Sabatini, J., Lee, Y., & Ventura, M. (2007). The automated text adaptation tool. In B. Carpenter, A. Stent, & J. D. Williams (Eds.). *Proceedings of Human Language Technologies* (pp. 3–4). The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations.
- Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Companion volume, Council of Europe Publishing. www.coe.int/lang-cefr.
- Dragićević, R. (2018). *Srpska leksika u prošlosti i danas.* Matica srpska.
- Krajišnik, V. (2016). *Leksički pristup srpskom kao stranom jeziku. Edicija Jezik, književnost, kultura, knjiga 8.* Univerzitet u Beogradu, Filološki fakultet.

- Laposhina, A. N., & Lebedeva, M. Y. (2021). Tintometr: an online tool for automated complexity level assessment of texts for Russian language learners. *Russian Language Studies*, *19*(3), 331–345.
- Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). The Seventh Workshop on Balto-Slavic Natural Language Processing. Association for Computational Linguistics.
- Medenica, L. (2023). Uloga stepenovane lektire i leksičkih minimuma u nastavi i učenju srpskog kao stranog jezika na nivou A1. U V. Krajišnik (ur.). *Srpski kao strani jezik u teroriji i praksi V: tematski zbornik radova* (pp. 337–347). Međunarodni naučni skup Srpski kao strani jezik u teoriji i praksi, 28. i 29. oktobar 2022. Filološki fakultet ‑ Centar za srpski kao strani jezik.
- Stanković, R., Šandrih, B., Krstev, C., Utvić, M., & Škorić, M. (2020). Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, D. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.). *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3954–3962). European Language Resources Association.
- Stanković, R., & Škorić, M. (2021). *SrpKor4Tagging-TreeTagger* (Version 1.0.0) [Model]. https://doi.org/10.57771/bvkk-jv85.
- Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. ArXiv preprint.* https://arxiv.org/abs/2308.04255

Summary

Programs for automatic text correction and adaptation have an increasing influence on modern language didactics. Given that such tools are not used in teaching Serbian as a foreign language, the goal of our research was to determine the lexical minimums for level A1 in Serbian as a foreign language and create an algorithm for their detection, i.e., to take the first steps in developing a program for automatic text correction and adaptation. In the research, we used a methodology that relies on the linguo-didactics of Russian as a foreign language, taking into account the developed corpora of lexical minimums in that Slavic language and the existence of the online tool Textometr for checking the complexity of the text. In this way, we identified 783 lexemes that represent the list of lexical minimums for level A1 in Serbian as a foreign language. Then, for the purposes of the paper, we created an algorithm in the Python programming language, which we tested on a specific text and found certain shortcomings when it comes to lemmatization of the text. In the following period, it is necessary to refine the lemmatizer in order to develop a program for automatic text correction.

***Keywords**: Serbian language as a foreign language, level A1, lexical minimum, automatic correction and text adaptation*