811.163.41'322.2
811.133.1'322.2
https://doi.org/10.18485/msc.2022.51.1.ch12

Душица В. ТЕРЗИЋ*
Универзитет у Београду
Филолошки факултет

Оригинални научни рад
Примљен: 11. 2. 2022.
Прихваћен: 2. 3. 2022.

СРПСКИ ЈЕЗИК У КОНТЕКСТУ ТЕНДЕНЦИЈА У РАЧУНАРСКОЈ ОБРАДИ ЈЕЗИКА**

Утицај технологије на језик може се прилагођавати и контролисати кроз развој језичких ресурса и алата у оквирима рачунарске обраде језика. У овоме раду представљамо начине рачунарске обраде текста, као и рачунарске методе у основи описаних поступака. Затим приказујемо постојеће ресурсе и алате за рачунарску обраду српскога језика. Указујемо на неповољан положај српскога језика у овоме контексту, који је последица слабе развијености ресурса, нарочито за рачунарску анализу текстова на семантичком нивоу. На основу анализе постојећих ресурса, алата те тенденција у рачунарској обради других језика, излажемо модел израде лексичких мрежа на основу постојећег паралелног корпуса. Циљ овога модела је богаћење ресурса за семантичку анализу текстова на српскоме језику, чиме би се како олакшала лингвистичка семантичка истраживања, тако и утицало на повољнији положај српскога језика у контексту језичких технологија.

Кључне речи: језичке технологије, рачунарска обрада језика, семантичко означавање, лексичке мреже, паралелни корпуси, српски језик, француски језик.

1. Увод

М. Тадић (2003: 9) сликовито пореди језичке технологије с бродом који плови немирним водама између две сигурне научне обале, лингвистике и информатике. Рачунари су као неизоставни део сваког сегмента људскога деловања данас пронашли место и у лингвистичким истраживањима. С друге стране, лингвистичка су знања незаобилазна при изради успешних рачунар-

* dusica.terzic@fil.bg.ac.rs

** Рад је настао у склопу билатералног француско-српског пројекта *PREDICT: Мали корпусно заснован француско-српски речник предлога* (*PREDICT : Petit dictionnaire raisonné des prépositions : français-serbe*) који подржава Министарство просвете, науке и технолошког развоја Републике Србије и Министарство за Европу и спољне послове те Министарство високог образовања, науке и иновација Републике Француске у оквиру програма „Павле Савић" (партнерство Hubert Curien) за 2020. и 2021. годину.

ских алата који на било који начин обрађују текст. Овај рад сврстава се у оквире рачунарске обраде језика, младе, нестабилне, али значајне области на размеђи информатике и лингвистике.

Према С. Фабр (2012: 1), сврха данашњих система рачунарске обраде језика налази се у могућности анализе огромне количине текста и у примени те анализе у конкретним практичним поступцима, као што су екстракција података, класификација докумената и претрага података. Као примери језичких технологија које користе ресурсе развијене у оквиру рачунарске обраде текста могуће је навести интернетске претраживаче, системе за исправљање правописних и граматичких грешака у тексту, програме који претварају текст у говор и говор у текст, те виртуелне асистенте. Будући да се програми засновани на језичким технологијама свакодневно користе у комуникацији, њихов је утицај на развој језика значајан. Међутим, утицај технологија на језик може се контролисати кроз пажљиву и планску израду ресурса и алата за рачунарску језичку обраду. За обраду српскога језика, пак, рачунарски су ресурси слабо развијени у поређењу с другим европским језицима (уп. Милетић 2018: 18). Стога је циљ анализе изложене у овоме раду да се – на основу прегледа принципа по којима системи за рачунарску обраду текста функционишу, као и на основу приказа начина на који они спроводе лингвистичку анализу, на основу предности и ограничења постојећих ресурса и алата – укаже на поступак којим би се положај српскога језика у домену рачунарске обраде побољшао.

Најпре, у одељку 2 описујемо на које се све начине текстови могу лингвистички анализирати помоћу рачунара. Затим у одељку 3 објашњавамо методе у основи језичких алата. Анализирамо потом у одељку 4 већ постојеће ресурсе развијене за рачунарску обраду српскога језика. Посебну пажњу посвећујемо семантичком нивоу обраде да бисмо у одељку 5 предложили поступке израде ресурса за семантичку обраду српскога језика.

2. Нивои рачунарске анализе текста

Да би рачунар могао анализирати неки текст, неопходно је означити од којих се делова дати текст састоји. Стога се у рачунарско-лингвистичкој анализи најпре означава логичка структура текста, што подразумева означавање почетка и краја одређеног дела текста који је машински читљив, односно који рачунар може разумети (Васиљевић 2014: 4). Подела текста може се вршити на нивоу поглавља, пасуса или реченица. Овај се поступак назива *сегментирањем* (Васиљевић 2014: 11). Подела корпусног текста на појавнице најчешће се назива токенизацијом. У појавнице се убрајају како речи из текста, тако и сви интерпункцијски знаци и цифре (Ерјавец и др. 2015: 263).

Када се текст сегментира и токенизује, чиниоцима тако означене логичке структуре могу се додати подаци који описују текст на неком од лингвистичких нивоа поступком који се назива *означавање* (Тадић 2003: 32). Тако можемо говорити о фонетском, лематском, морфолошком, синтаксичком, се-

мантичком, дискурзивном те прагматичком означавању. Означавање на једном нивоу олакшано је уколико је текст већ означен на неком другом нивоу, али се ипак у нивоима означавања може уочити хијерархија према сложености самога поступка и према развијености алата и ресурса на датом нивоу (Ивон 2007: 5–7). Будући да у литератури није било речи о фонетском, дискурзивном и прагматичком означавању српскога језика, говоримо само о четири типа означавања за које смо у литератури о рачунарској обради српског језика пронашли довољно података.

Прво, морфосинтаксичко означавање подразумева одређивање не само врсте речи, већ и информација о њиховим морфосинтаксичким категоријама (Тадић 2003: 32). Овај поступак подразумева да се појавницама додели ознака с одговарајућом морфосинтаксичком информацијом из претходно дефинисаног затвореног скупа. Друго, лематско означавање је процес приликом кога се свакој појавници додаје податак о њеној леми или канонском облику (Агић и др. 2013: 48). Као канонски облик глагола означава се инфинитив, код именица номинатив једнине, код придева номинатив мушког рода једнине позитива, код заменица прво лице једнине мушког рода. Треће, у процесу синтаксичке анализе или парсирања обележавају се синтаксички односи између две појавнице у одређеној реченици, односно функција речи у реченици. Алати којима се текст може парсирати називају се парсерима. М. Тадић (2003: 33) парсере дели на плитке парсере који анализирају односе зависности међу деловима реченице, дубоке парсере који спроводе пуну синтаксичку анализу до нивоа речи, те робусне парсере који анализирају све комбинације реченичних делова, чак и оне које нису граматичне (робусни парсери). Производ парсирања је синтаксичко стабло, а синтаксички означен корпус назива се банком стабала (енг. *treebank*) (Тадић 2003: 32; Утвић 2013: 111).

Четврто, семантички се појавнице могу означити на више начина. Могу се означити и класификовати именовани ентитети (властите именице, називи компанија, датуми и сл.), као и стручни термини у некоме тексту (Ерјавец и др. 2015: 264). Поред тога, постоје системи који појавнице повезују с појмовима у неком спољашњем ресурсу или им додељују семантичку улогу (Ерјавец и др. 2015: 264). Пример таквог спољашњег ресурса су лексичке мреже као што је је *WordNet*. Ради се о електронској лексичкој бази у којој су речи и групе речи повезане лексичким и појмовним везама у виду означених лукова и груписане су у скупове синонима, тзв. синсетова (енг. *synsets*) (Фелбаум 2006: 665). На пример, реч „добар" је повезана антонимском везом с речју „лош", која је даље синонимском везом повезана с речју „покварен". Део лексичке мреже који није својствен одређеном језику назива се међујезичким индексом (Interlingual Index – ILI) (Фелбаум 2006: 669). Путем овога индекса повезују се појмови сличнога значења у различитим језицима. Тиме се олакшава употреба лексичких мрежа у вишејезичним апликацијама, али отежава њихово одржавање будући да их је потребно ажурирати са сваким новим издањем неке од лексичких мрежа (Младеновић и др. 2014: 52).

Приликом означавања на свакоме од ових нивоа морају се донети разноврсне одлуке које ће утицати на сваки каснији корак, али и на употребљивост обрађених текстова у пракси. Једна од одлука које је потребно донети при, на пример, морфосинтаксичком означавању тиче се броја морфосинтаксичких ознака у скупу. На ову одлуку првенствено утиче природа језика који се обрађује. Код језика сложене флективне морфологије, као што је српски, значајно је одредити све морфолошке категорије. За такве језике скуп ознака је нужно бројан. Примера ради, скуп морфосинтаксичких ознака које А. Милетић користи при означавању српског садржи 1042 ознаке, док се при означавању енглеског корпуса PennTreeBank користи 36 ознака (Милетић 2018: 55, 120). Уколико користимо већи број ознака, успешност означавања је мања (Агић и др. 2013: 52). С друге стране, уколико су означене не само врсте речи, већ и њихова детаљна морфосинтаксичка својства, каснији поступак синтаксичког означавања је прецизнији будући да су подаци о синтаксичкој функцији неретко садржани у морфосинтаксичким одликама (Милетић и др. 2016: 507). Слични компромиси се морају правити и када се бира одговарајући програм за рачунарску обраду језика. Приступи на којима су засновани ови програми описани су у одељку 3.

3. Приступи у основи програма за рачунарску обраду текста

У основи програма за рачунарску обраду текста стоје два супротстављена приступа, лингвистички и рачунарски. Када се овој области приступа из лингистичке перспективе, циљ је описати језик у целини што прецизније. Рачунарски приступ тежи обради што већег броја језичких података у што краћем времену, иако се тиме, можда, нарушава прецизност (Тадић 2003: 9–11). Први приступ је у основи система заснованих на низу правила, док други одговара системима у чијој су основи статистички модели.

Правила на којима је заснована прва врста система представљају упутства како да се обради одређена појавница или структура, коју рачунару даје истраживач, а која су најчешће груписана у тзв. формалне граматике (Абеје 1990). Премда је предност ових система у тежњи ка прецизности, свака језичка структура која се јави у тексту који треба обрадити, а није описана правилима или није обухваћена том формалном граматиком, не може бити обрађена у програмима који су на тим правилима засновани (уп. Ђорђевић 2017: 3–4). Уз то, ова су правила подвргнута опажањима појединца или групе појединаца који би ту граматику и правила саставили.

Други приступ у основи програма за рачунарску обраду језика почива на статистичким системима који су најчешће засновани на моделима надгледаног машинског учења. Надгледано машинско учење је процес приликом кога се најпре скуп текстова означи ручно – или аутоматски, а ознаке се провере ручно – а затим софтвер анализира ручно означен скуп текстова који се

назива „корпус за учење" и тако ствара модел на основу кога „учи" како да означи неки нови сличан неозначени корпус (Уријели 2013: 16). Предност статистичких програма је обрада сваке језичке структуре која се јави у тексту, премда нису увек прецизни (Ерјавец и др. 2015: 264). Поред тога, ови се системи увежбавају на корпусима који представљају реалну слику језика. Тиме се покушава избећи субјективност, која је при састављању статистичких модела умногоме смањена у односу на израду правила и формалне граматике (в. горе). Ограничења која се у овоме погледу могу имати тичу се природе корпуса за учење.

Ниједан од ова два система није савршен, те при одабиру система за рачунарско-лингвистичку анализу треба водити рачуна о томе који је примарни циљ анализе. Уколико је прецизност важнија од брзине и покривености структура, онда предност имају системи засновани на правилима. Уколико је, пак, за анализу битно не само обрадити сваку појавницу у тексту, већ и обрадити је брзо науштрб прецизности, бира се систем надгледаног машинског учења. Истраживања (нпр. Ерјавец и др. 2015: 271) показују да статистички системи постижу боље резултате од система заснованих на правилима или граматикама. Ипак, најбоље смо резултате у литератури нашли за системе који могу интегрисати како статистику, тако и правила (уп. Уријели 2013). Без обзира на то који је принцип у основи рачунарског алата, сви они при обради текста користе језичке ресурсе, о којима пишемо у наставку.

4. Ресурси за рачунарско означавање српскога језика

М. Тадић у језичке ресурсе убраја лексиконе и корпусе (Тадић 2003: 31). Када говоримо о лексиконима у контексту рачунарско-лингвистичке анализе текста, мислимо на машински читљиве речнике за којима посежу рачунарски програми при обради текста (Тадић 2003: 27; Ђорђевић 2017: 6). Пример ових речника су морфолошки лексикони који садрже флективне облике и ознаке морфосинтаксичких својстава ових облика (Бенко/Мур 2004). Иако су системи који се служе овим речницима изузетно успешни, израда морфолошког речника за језике са сложеном флективном морфологијом дуготрајан је и скуп процес.

Ни израда корпуса за рачунарско-лингвистичку анализу није једноставан ни јефтин поступак. Корпуси су нарочито значајни за обраду помоћу статистичких рачунарских програма. Као што смо већ описали (в. одељак 3), они се могу користити за „учење" статистичког модела који ће бити примењен на друге корпусе. Променом корпуса за учење, један те исти програм може се прилагодити новоме језику. Развој статистичких система умногоме, дакле, зависи од постојања означених корпуса за учење, као и од врсте текстова похрањених у њима. Иако литература (нпр. Крстев и др. 2004б; Утвић 2013; Ђорђевић 2017) пружа увид у значајан број ресурса и алата развијених за обраду српскога језика, они се често не могу преузети и користити у даљим истраживањима. У наставку, стога, при опису ресурса за српски језик

као значајан критеријум издвајамо доступност. Како израда ресурса захтева тимски рад, представљамо ресурсе према томе која их је група истраживача развила.

Група истраживача с Математичког факултета Универзитета у Београду израдила је први означени копус српскога језика који се састоји од превода романа *1984* Џорџа Орвела упареног с енглеским изворником и сегментираног, токенизованог и означеног морфосинтаксички и лематски (Крстев и др. 2004а). На основу њега је израђен лексикон који садржи све флективне облике из корпуса којима је придружена лема и детаљна морфосинтаксичка ознака (Крстев и др. 2004а). Оба се ресурса могу слободно преузети.[1] Међу ресурсима који су даље развијани у оквиру ове групе, а који су доступни на интернетској страници[2] Математичког факултета, три корпуса су доступна за претрагу. Прво, *SrpKor2013* (Утвић 2011) корпус је савременог српског језика из различитих домена (113 милиона појавница) с библиографским и лематским ознакама и ознаком врсте речи. Друго, доступна су два паралелна корпуса, *SrpEngKor* и *SrpFranKor*. Српско-енглески корпус *SrpEngKor* од 4,4 милиона појавница садржи правне, књижевне, новинске текстове, те титлове филмова (Крстев/Витас 2011), док је *SrpFranKor* српско-француски корпус од 1,7 милиона појавница из књижевних и новинских текстова (Крстев/Витас 2006).

У оквиру сарадње Универзитета Тулуз – Жан Жорес с Филолошким факултетом Универзитета у Београду састављен је паралелни француско-српско-енглеско-шпански корпус *ParCoLab*[3] (Милетић и др. 2017; Марјановић и др. 2018а), који тренутно садржи текстове с преко 32,9 милиона појавница (Терзић и др. 2020: 67), чија се база непрекидно допуњује текстовима разноврсних жанрова. Ресурси за рачунарску обраду језика који су развијени у оквиру ове групе су: *wikimorph-sr*, морфолошки лексикон српског језика од 1.226.638 облика (Милетић 2017); *ParCoTrain*, корпус за учење и евалуацију морфосинтаксичког и лематског означавања српског језика од око 150.000 појавница (Милетић 2013); *ParCoTrain-Synt*, корпус књижевних текстова, чија прва верзија садржи око 80.000 појавница (Милетић 2018); *ParCoJour*, корпус новинских текстова од 30.000 појавница (Терзић 2019; Терзић 2020). Оба корпуса су морфосинтаксички, лематски и синтаксички означена системима надгледаног машинског учења. ParCoLab се може претраживати бесплатно преко сучеља за претрагу без писања молбе за приступ, док се сви наведени ресурси могу слободно преузети.

Значајан број доступних ресурса развија и група хрватских истраживача (в. Агић и др. 2013; Љубешић и др. 2016)[4], међу којима издвајамо лематски, морфолошки и синтаксички означен корпус *Setimes.HR* од 90.000 појавница

[1] Може се преузети с адресе https://www.clarin.si/repository/xmlui/handle/11356/1043. Овој и свим осталим везама последњи пут смо приступили 2. фебруара 2020.

[2] http://www.korpus.matf.bg.ac.rs/prezentacija/korpusi.html.

[3] http://parcolab.univ-tlse2.fr/en/about/resources/.

[4] Могу се пронаћи на следећој адреси: https://github.com/ffnlp/sethr и https://www.clarin.si/repository/xmlui/.

(Агић/Љубешић 2015), за који је развијена и српска верзија[5] (Самарџић и др. 2017).

Литература пружа увид у велики број експеримената у којима су вредноване успешности разноврсних алата у процесу морфосинтаксичког, лематског и синтаксичког означавања користећи ресурсе за српски језик (Агић и др. 2013; Гесмундо/Самарџић 2012; Љубешић и др. 2016; Милетић 2018; Терзић 2019; Утвић 2011). М. Утвић (2011: 47) анализира постојеће алате за морфосинтаксичко означавање користећи корпус *SrpKor2013*. Након анализе постојећих алата, предност даје статистичким програмима, који достижу изузетно високу прецизност: она у просеку износи 96,57%. Међутим, скуп морфосинтаксичких ознака за означавање корпуса *SrpKor2013* сведен је на 16 у првим експериментима. С друге стране, А. Милетић (2018) користећи 1.042 ознаке при означавању корпуса *ParCoTrain-Synt* достиже прецизност од 85%. Што се тиче лематског означавања, највише резултате (97,72%) достижу А. Гесмундо и Т. Самарџић (2012) при означавању корпуса *1984*. Међутим, незнатно ниже резултате (редом 96,3% и 96,5%) достижу Ж. Агић и др. (2013) при означавању корпуса *SETimes* и А. Милетић (2018) при означавању корпуса *ParCoTrain-Synt* у много краћем временском року.

Формалну граматику за синтаксичку анализу српског језика под називом *SrpTAG* развила је Б. Ђорђевић (Ђорђевић 2017). Ова граматика се користи за аутоматску синтаксичку анализу реченица. Међутим, она за сада препознаје само оне реченице у којима су реченични аргументи у канонском редоследу, односно јављају се овим редом: субјекат, предикат, објекат. Стога овај алат многе иначе граматичне реченице означава као аграматичне. Прави и неправи објекат тренутно препознаје само уколико се налазе иза глагола, а граматички субјекат уколико се налази на првом месту у реченици (Ђорђевић 2017: 204–205). Не препознају се футур у облику *да*+презент, модални глаголи, предлошко-падежне конструкције у упитном облику, као ни предикативи, енклитички облик логичког субјекта, те негација (Ђорђевић 2017: 206–207). Закључујемо да текст на српском језику не може бити значајно анализиран на синтаксичком нивоу помоћу ове граматике будући да она не препознаје неке учестале реченичне моделе. С друге стране, статистички алат *Talismane*, који користи А. Милетић (2018) за синтаксичку анализу српског језика, може анализирати сваку реченицу неозначеног корпуса српског језика. За увежбавање алата коришћен је корпус *ParCoTrain-Synt*, означен морфосинтаксички, лематски и синтаксички. Корпус је на крају вредноване мерама означеног придруживања (LAS) и неозначеног придруживања (UAS) и достигао је највише резултате парсирања које смо пронашли у литератури – UAS: 91,22 и LAS 87,48 (Милетић 2018). Међутим, сви су текстови у овоме корпусу књижевни, што потенцијално може представљати проблем за даљи развој ресурса.

[5] Доступна на адреси: https://universaldependencies.org/.

Из прегледа литературе и на основу доступности ресурса и алата за рачунарску обраду српскога језика може се закључити да за српски језик постоје доступни ресурси за означавање на лематском, синтаксичком и морфосинтаксичом нивоу. Следећи ниво у хијерархији односи се на семантичко означавање, коме у наредном делу посвећујемо посебну пажњу.

5. Семантичко означавање српскога језика

Као ресурс за семантичко означавање српскога језика, у литератури се наводи семантичко-лексичка мрежа која се назива српски *WordNet*. Рад на српској лексичкој мрежи започет је у оквиру пројекта балканске лексичке мреже *BalkaNet project*, који је заснован на моделу европске лексичке мреже *EuroWordNet* (EWN) (Крстев и др. 2004б: 147–148; Митровић и др. 2014: 52–53). Првобитно је донета одлука да се балканске лексичке мреже израде превођењем с енглеског, француског и других европских језика укључених у пројекат базичних концепата из EWN. Но, током рада одустало се од овога приступа будући да су се у оквиру пројекта уочили недостаци једнојезичке мреже за енглески језик при примени на друге језике (Крстев и др. 2004б: 147–148). Одлучено је да се прати развој принстонске лексичке мреже (PWN) (Фелбаум 2006) и да се концепти путем ILI индекса повежу с другом верзијом издања PWN (Крстев и др. 2004б: 148). Српска је лексичка мрежа имала 7.000 синсетова на крају пројекта у 2004. години, док је у наредним годинама обогаћена с још 14.000 синсетова из специфичних домена (емоције, биологија, биомедицина, религија, право, лингвистика, књижевност, библиотекарство, рачунарство и кулинарство) (Младеновић и др. 2014: 52–53), а додати су и појмови карактеристични за балканско подручје (Станковић и др. 2018: 104). Ручна израда синсетова, међутим, премда омогућава највећу прецизност, изискује толику количину људских ресурса и времена да многи започети пројекти стагнирају (Станковић и др. 2018: 104; Саго/Фишер 2008: 14).

Аутоматизовање превођења концепата из PWN убрзало би процес израде и проширења српске лексичке мреже. Међутим, аутори наводе да мане овога приступа произилазе из разлика у енглеском и српском језику (Крстев и др. 2004б: 149–150). Као пример наводи се да се енглеска именица *peer* (срп. 'парњак') често преводи придевом *раван* („Он је њему раван."), што је проблематично за систем, као што је *WordNet* (Крстев и др. 2004б: 150), у коме су синсетови класификовани према врсти речи. Стога је неопходно све овако генерисане синсетове проверити ручно. У овоме су процесу коришћени енглеско-српски и француско-српски корпуси развијени на Математичком факултету у Београду. Ту се види да су вишејезични паралелни корпуси значајно средство у процесу провере ваљаности синсета, као и средство откривања нових семантичких парова којима се лексичка мрежа може проширити (Крстев и др. 2004б: 152–156). Један од проблема претходних приступа лежи у неактуелности ресурса коришћених у процесу проширења. Наиме, ради се

о корпусима претежно књижевних текстова и текстова међу којима нема довољно савременог српског језика. Ресурсу смо успели приступити преко сучеља за претрагу[6], али одговор на молбу за његово преузимање нисмо добили до завршетка писања овога чланка, иако аутори наводе да је могуће преузети овај ресурс за некомерцијалну употребу (Младеновић и др. 2014: 53).

Можемо, дакле, закључити, да за семантичко означавање српскога језика немамо на располагању одговарајући ресурс, те да би следећи корак било развијање доступне лексичке мреже. С истим су се проблемом сусрели истраживачи при изради француске лексичке мреже *WOLF* (Саго/Фишер 2008: 14). Њихова метода која олакшава израду лексичке мреже комбинује употребу паралелних корпуса који садрже француске текстове и двојезичне француско-енглеске ресурсе (Саго/Фишер 2008: 16). Евалуација је показала да је новостворена лексичка мрежа кориснија од претходно створених ресурса за синтаксичко означавање, премда, као и све претходне, има недостатке (Саго/Фишер 2008: 18).

На основу свега изложеног, закључујемо да би најефикаснији начин за израду нове српске лексичке мреже, или проширења већ постојеће, уколико бисмо добили приступ њој, била примена метода који су употребили истраживачи при изради француске лексичке мреже, а на основу паралелног корпуса *ParCoLab* као полазне тачке. Најпре, овај корпус је осим књижевним текстовима обогаћен и свакодневним говором уношењем транскрипата савремених играних и цртаних филмова и њихових превода, како на преведеном, тако и на изворном српском језику (Терзић и др. 2020). Затим, *ParCoLab* садржи реченице из четири језика (француског, енглеског, српског и шпанског). Упарене реченице ових језика ручно су упарене. Будући да енглеска лексичка мрежа служи као полазна тачка за израду лексичких мрежа за остале језике, а да је израда француске мреже на описани начин већ дала боље резултате (в. Саго/Фишер 2008: 15), аналогно би и синсетови направљени на овај начин, очекујемо, били прецизнији.

6. Закључак

У овоме раду представљен је српски језик у контексту рачунарске обраде језика. Објашњени су нивои на којима су текстови у српскоме језику досад означавани, односно морфосинтаксичко, лематско и синтаксичко означавање. Затим су описани принципи који се налазе у основи рачунарских система за обраду српскога језика. Анализа доступних резултата показује да су статистички системи у које се могу интегрисати и правила најефикаснији системи за означавање текстова на српскоме језику.

[6] http://dcl.bas.bg/bulnet/. Последњи приступ 6. фебруара 2022.

За њихово функционисање потребни су поуздани, и то обимни ресурси. Ресурси за обраду српскога језика у литератури су релативно бројни, али су ретко доступни за слободну употребу, па их је тешко самостално проверити и применити. Будући да је њихова слободна дистрибуција предуслов за плодоносније развијање ресурса и побољшање статуса српског језика у домену рачунарске обраде, осврнули смо се на оне ресурсе који се могу слободно преузети. Уочили смо да је тренутно највећа потреба за доступном семантичко-лексичком мрежом којом би се текстови могли семантички обрађивати. Српски *WordNet* је потанко описан у литератури, али није могуће добити приступ за преузимање. Стога смо закључили да је следећи практичан корак израда лексичке мреже по моделу израде француске лексичке мреже *WOLF*. Ова је мрежа израђена употребом паралелних корпуса и двојезичних интернетских речника. За то је на располагању паралелни корпус *ParCoLab* који, поред текстова на српском језику, садржи и разноврсне савремене текстове на француском и енглеском језику.

ЛИТЕРАТУРА

Абеје 1990: A. Abeillé, Lexical and Syntactic Rules in a Tree Adjoining Grammar, *u*: *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, Pittsburgh: Association for Computational Linguistics, 292–298.

Агић и др. 2013: Ž. Agić i dr., Lemmatization and morphosyntactic tagging of Croatian and Serbian, *u*: *Proceedings of The 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, Sofija: Association for Computational Linguistics, 48–57.

Агић/Љубешић 2015: Ž. Agić, N. Ljubešić, Universal Dependencies for Croatian (that work for Serbian, too), *u: Proceedings of The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, Hissar: Association for Computational Linguistics (ACL), 1–8.

Бенко/Мур 2004: M. Banko, R. Moore, Part of speech tagging in context, *u: COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, Morristown: Association for Computational Linguistics, 556–562.

Васиљевић 2014: Н. Васиљевић, *Аутоматска обрада правних текстова на српском језику* (необјављена докторска дисертација, Београд: Филолошки факултет).

Гесмундо/Самарџић 2012: A. Gesmundo, T. Samardžić, Lemmatising Serbian as category tagging with bidirectional sequence classification, *u*: *Proceedings of The 8th Language Resources and Evaluation Conference (LREC 2012)*, Istanbul: European Language Resources Association, 2103–2106.

Ђорђевић 2017: B. Đorđević, *Izrada osnova formalne gramatike srpskog jezika upotrebom metagramatike* (doktorska disertacija, Beograd: Filološki fakultet Univerziteta u Beogradu).

Ерјавец и др. 2015: T. Erjavec i dr., Jezikovne tehnologije in zapis korpusa, *u*: *Slovar sodobne slovenščine: Problemi in rešitve*, Ljubljana: Znanstvena založba Filozofske fakultete Univerziteta v Ljubljani, 262–276.

Ивон 2007: Yvon, F. *Une petite introduction au traitement automatique des Langues Naturelles.* ‹https://www.researchgate.net/publication/228545119_Une_petite_introduction_au_Traitement_Automatique_des_Langues_Naturelles›.02.02.2022.

Крстев и др. 2004а: C. Krstev i dr., MULTEXT-East resources for Serbian, *u*: *Proceedings B of the 7th international multiconference information society:* Language technologies, Ljubljana: Jožef Stefan Institute, 108–114.

Крстев и др. 2004б: C. Krstev i dr., Using Textual and Lexical Resources in Developing Serbian Wordnet, Bucharest: *Romanian Journal of Information Science and Technology*, 7(1–2), 147–161.

Крстев/Витас 2006: C. Krstev, D. Vitas, Literature and aligned texts, *u*: M. Slavcheva, i dr. (eds) *Readings in Multilinguality*, Sofia: Institute for Parallel Processing, Bulgarian Academy of Sciences, 148–155.

Крстев/Витас 2011: C. Krstev, D. Vitas, An aligned English-Serbian corpus, *u*: N. Tomović, N., J. Vujić, (eds), *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality)*, vol. 1, Belgrade: Faculty of Philology, 495–508.

Љубешић и др. 2016: N. Ljubešić i dr., New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian, *u*: *Proceedings of The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož: European Language Resources Association (ELRA), 4264–4270.

Марјановић и др. 2018а: С. Марјановић и др., Паралелни корпус *PARCOLAB* у служби српско-француске лексикографије, *Српско-француске књижевне и културне везе у европском контексту*, (ред. Ј. Новаковић, М. Сребро), Нови Сад: Матица српска.

Марјановић и др. 2018б: S. Marjanović i dr., A sample French-Serbian Dictionary Entry based on the ParCoLab Parallel Corpus, *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana: Ljubljana University Press, Faculty of Arts, 423–435.

Младеновић и др. 2014: M. Mladenović i dr., Developing and maintaining a wordnet: Procedures and tools, *Proceedings of the 7th International Global WordNet Conference*, (eds H. Orav i dr.), Tartu: University of Tartu Press, 55–62.

Милетић 2013: A. Miletic, *Annotation morphosyntaxique semi-automatique d'un corpus littéraire serbe*, Mémoire de master, Lille: Université Charles de Gaulle – Lille 3.

Милетић и др. 2016: A. Miletic i dr., Mise au point d'une méthode d'annotation morphosyntaxique fine du serbe, *Actes de Traitement Automatique des Langues Naturelles (TALN 2016)*, Paris: AFCP – ATALA, 506–514.

Милетић 2017: A. Miletic, Building a morphosyntactic lexicon for Serbian using Wiktionary, *Actes de Sixièmes Journées d'études Toulousaines (JéTou2017) : Les interfaces en sciences du langage*. Toulouse, 30–34.

Милетић и др 2017: A. Miletic i dr., ParCoLab: A Parallel Corpus for Serbian, French and English, *Text, Speech and Dialogue TSD 2017*, (eds K. Ekštein, V. Matoušek), Lecture Notes in Computer Science, vol. 10415, Cham: Springer, 156–164.

Милетић 2018: A. Miletic, *Un treebank pour le serbe : constitution et exploitations*, thèse de doctorat, Toulouse: Université Toulouse – Jean Jaurès.

Саго/Фишер 2008: B. Sagot, D. Fišer, Building a free French wordnet from multilingual resources, *u*: *Proceedings of OntoLex 2008*, Marrakech, 14–19.

Самарџић и др 2017: T. Samardžić i dr., Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages, *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, (eds T. Erjavec i dr.), Valencia: Association for Computational Linguistics, 39–44

Станковић и др. 2018: R. Stanković i dr., Resource-based WordNet Augmentation and Enrichment, *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018)*, Sofia: The Institute for Bulgarian Language, Bulgarian Academy of Sciences, 104–114.

Тадић 2003: M. Tadić, *Jezične tehnologije i hrvatski jezik*, Zagreb*:* Ex Libris.

Терзић 2019: D. Terzic, Parsing des textes journalistiques en serbe par le logiciel Talismane, *Actes de la conférence TALN-RECITAL*, (eds A.-L. Ligozat, S. Ghannay) *(Conférence sur le Traitement Automatique des Langues Naturelles) PFIA 2019*, Toulouse: AfIA, 591–604.

Терзић 2020: Д. Терзић, Анализа успешности парсера *Talismane* на нивоу синтаксичких етикета у корпусу ParCoJour, *Језици и културе у времену и простору, 9.2*, (ред С. Гудурић), Нови Сад: Филозофски факултет, 265–279.

Терзић и др. 2020: D. Terzić i dr., Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data*Text, Speech, and Dialogue, TSD 2020*, (eds P. Sojka i dr.), Lecture Notes in Computer Science, vol 12284, Cham: Springer, 61–70, https://doi.org/10.1007/978-3-030-58323-1_6

Уријели 2013: A. Urieli, *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*, PhD thesis, Toulouse: Université Toulouse le Mirail – Toulouse II.

Утвић 2011: М. Утвић, Анотација корпуса савременог српског језика. *Инфотека* 12 (2), 2011, 39–51. ‹http://infoteka.bg.ac.rs/index.php/sr/arhiva/2011/2/infoteka-12-2-2011-39-51›.02.02.2022.

Утвић 2013: M. Utvić, *Izgradnja referentnog korpusa savremenog srpskog jezika* (neobjavljena doktorska disertacija), Beograd: Filološki fakultet.

Фабр 2012: C. Fabre, Traitement automatique de textes : techniques linguistiques. *Encyclopédie des techniques de l'ingénieur – Représentation et traitement des documents numériques.* Editions Techniques de l'Ingénieur.

Фелбаум 2006: C. Fellbaum, WordNet(s), *Encyclopedia of Language & Linguistics*, Second Edition, volume 13, (ed K. Brown), Oxford: Elsevier, 665–670.

Dušica Terzić

LA LANGUE SERBE DANS LE CONTEXTE DU TRAITEMENT AUTOMATIQUE DES LANGUES

Résumé

Le présent article illustre la position de la langue serbe dans le domaine du traitement automatique des langues naturelles (TAL). Nous avons défini d'abord la segmentation, la tokénisation, l'étiquetage morphosyntaxique et sémantique, la lemmatisation et le parsing dans le contexte des traits distinctifs du serbe. À savoir, il s'agit d'une langue à morphologie flexionnelle riche, ce qui rend difficile l'automatisation de ces tâches. Nous avons ensuite présenté les modèles traditionnellement utilisés dans le traitement du serbe, en soulignant que les meilleurs résultats présentés dans les travaux sont atteints en utilisant des systèmes statistiques qui peuvent aussi intégrer des règles heuristiques. Étant donné que les systèmes statistiques requièrent des ressources volumineuses et bien rédigées pour fonctionner efficacement, nous avons présenté les ressources développées pour le traitement du serbe en insistant sur leur libre disposition. Il s'avère qu'un grand nombre de ressources décrites dans la littérature n'est pas librement disponible. C'est surtout le cas du traitement au niveau sémantique dû au fait que le *WordNet* développé pour le serbe n'est pas mis en disposition. Par conséquent, nous consacrons toute une partie de l'article à la proposition du développement et enrichissement du *WordNet* serbe. Nous proposons la méthode utilisée pour la création du réseau sémantique français *Wolf* qui combine l'utilisation des corpus multilingues et des dictionnaires électroniques bilingues vu qu'un corpus français-anglais-serbe-espagnol *ParCoLab* se trouve à notre disposition.

Mots clés: technologies du langage, traitement automatique des langues, étiquetage sémantique, wordnets, corpus parallèles, serbe, français.