Otvoreni resursi i tehnologije za obradu srpskog jezika

Otvoreni resursi i tehnologije za obradu srpskog jezika

Vuk Batanović[1], Nikola Ljubešić[2], Tanja Samardžić[3], Maja Miličević Petrović[4]

[1]*Inovacioni centar Elektrotehničkog fakulteta, Beograd, Srbija*
[2]*Institut Jožef Stefan, Ljubljana, Slovenija*
[3]*Univerzitet u Cirihu – URPP "Language and Space", Cirih, Švajcarska*
[4]*Univerzitet u Beogradu – Filološki fakultet, Beograd, Srbija*
vuk.batanovic@ic.etf.bg.ac.rs, nikola.ljubesic@ijs.si, tanja.samardzic@uzh.ch, m.milicevic@fil.bg.ac.rs

Rezime: Otvorenost jezičkih resursa i alata je od velike važnosti za povećanje kvaliteta i brzine razvoja tehnologija za računarsku obradu prirodnih jezika. U ovom radu predstavljeni su otvoreni resursi za obradu srpskog jezika. Opisani su ručno anotirani korpusi, kao i širi spektar alata i računarskih modela, uključujući i veb servis koji omogućava njihovo jednostavno korišćenje.

Ključne reči: računarska lingvistika; korpusi tekstova; jezički alati; NLP; otvorena kultura.

I. Uvod

U poslednjih nekoliko godina došlo je do razvoja otvorenih, slobodno dostupnih resursa i tehnologija za računarsku obradu tekstova na srpskom jeziku, uključujući anotirane jezičke korpuse, alate za automatsku analizu i označavanje tekstova, kao i različite vrste modela za obradu prirodnih jezika (engl. *natural language processing* – NLP). Otvorenost i javna dostupnost omogućavaju veći stepen ponovljivosti rezultata NLP istraživanja, pospešuju saradnju istraživača i stimulišu zajedničko unapređivanje postojećih korpusa, alata i modela, umesto stalnog zasebnog pravljenja novih resursa istog tipa iz početka. Ova otvorenost je naročito važna u manjim jezicima poput srpskog, kod kojih se NLP istraživanjima bavi dosta ograničen krug naučnika i koji nisu od naročitog interesa za širu svetsku istraživačku zajednicu. Poseban doprinos promovisanju aktuelnosti i važnosti otvorenih jezičkih resursa za srpski i srodne jezike dao je projekat *Regional Linguistic Data Initiative* (ReLDI) [1], koji je doveo do nastanka većeg broja alata primenjivih na više južnoslovenskih jezika, kao i korpusa anotiranih korišćenjem standardizovane metodologije označavanja.

U ovom radu su ukratko prikazani aktuelni otvoreni, slobodno dostupni jezički resursi koji se mogu koristiti za analizu i obradu tekstova na srpskom jeziku. To uključuje kako resurse koji su proistekli iz ReLDI projekta, tako i one koje su autori rada razvili naknadno ili nezavisno od pomenutog projekta. Pri tome, pod otvorenim resursima ili alatima se podrazumevaju oni koji su javno dostupni na određenom repozitorijumu ili sajtu i jasno objavljeni pod nekom od odgovarajućih slobodnih licenci kao što su Creative Commons, GPL, i sl.

Ostatak rada je strukturiran na sledeći način: najpre su prikazani anotirani korpusi tekstova, a nakon toga alati i modeli za računarsku obradu tekstova na srpskom jeziku. Na kraju su izloženi planovi vezani za pravce daljeg razvoja.

II. Korpusi

U ovom odeljku su opisani ručno i automatski anotirani korpusi tekstova na srpskom jeziku koji su javno dostupni. Najpre su predstavljeni ručno anotirani korpusi standardnog i nestandardnog jezika, zatim korpusi posebno anotirani za određenu problematiku, dok je na kraju opisan veb korpus srpskog jezika. Izrada mnogih anotiranih korpusa je koordinisana sa sličnim poduhvatima na drugim jezicima, pre svega na hrvatskom i slovenačkom.

A. Ručno anotirani korpusi

SETimes.SR korpus [2], [3], izgrađen na osnovu *SETimes* paralelnog korpusa novinskih članaka [4], predstavlja ručno anotirani korpus tekstova pisanih standardnim srpskim jezikom namenjen obučavanju i evaluaciji računarskih modela na većem broju problema iz obrade prirodnih jezika. On sadrži 163 dokumenta podeljena na 3891 rečenicu, odnosno 86 726 tokena. *SETimes.SR* korpus je anotiran u pogledu segmentacije (na tokene, rečenice i dokumente), morfosintaktičkih oznaka, lema, sintaksnih dependencija kao i imenovanih entiteta. Korišćene morfosintaktičke oznake su u skladu sa MULTEXT-East v6 standardom za srpsko-hrvatski makrojezik[1]. Sintaksne dependencije su označene prema specifikaciji Universal Dependency v2 (UDv2)[2]. Oznake imenovanih entiteta su date u IOB2 formatu, uz razlikovanje pet tipova entiteta — osobe (PER), prisvojni pridevi izvedeni od imena (DERIV-PER), lokacije (LOC), organizacije (ORG) i razno (MISC). *SETimes.SR* je anotiran po uzoru na i uz pomoć modela obučenih na *SETimes.HR* [5] korpusu na hrvatskom jeziku, koji sada predstavlja deo većeg *hr500k* korpusa [6], [7].

ReLDI-NormTagNER-sr korpus [8], [9] je ručno označen skup tvitova napisanih na srpskom jeziku. Njegova primarna namena jeste prilagođavanje računarskih modela fenomenima koji su česti u nestandardnom jeziku koji se koristi na internetu. Korpus

[1] http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html

[2] http://universaldependencies.org

se sastoji od 3748 tvitova koji sadrže ukupno 91 781 token. Slojevi anotacije prisutni u korpusu su sledeći: tokenizacija i podela na rečenice, normalizacija na nivou reči, morfosintaktička anotacija (na osnovu standarda MULTEXT-East i Universal Dependencies), lematizacija i imenovani entiteti (sa istih 5 tipova entiteta kao i u *SETimes.SR* korpusu). Ovaj korpus je izgrađen uporedo sa sličnim resursom za hrvatski jezik – *ReLDI-NormTagNER-hr* [10] – a oba su oblikovana po uzoru na slovenački korpus *Janes-Tag* [11], [12].

B. Specijalizovani ručno anotirani korpusi

Na srpskom jeziku su javno dostupni ručno anotirani specijalizovani korpusi koji se odnose na semantičke probleme određivanja semantičke sličnosti, detekcije parafraza i analize sentimenta.

Za problematiku određivanja semantičke sličnosti kratkih tekstova izrađen je korpus *STS.news.sr* [13]. On sadrži 1192 para rečenica iz novinskog domena anotirana u pogledu stepena semantičke sličnosti između rečenica u paru. Ocene sličnosti su granulirane, na skali od 0 do 5, i dobijene su usrednjavanjem individualnih ocena petoro anotatora. *STS.news.sr* je kreiran korišćenjem sadržaja ranijeg korpusa parafraza *paraphrase.sr* [14], [15], koji sadrži samo binarne ocene sličnosti, ručno zadate od strane jednog anotatora, koje govore da li se rečenice u okviru para mogu smatrati parafrazama ili ne.

Za probleme analize sentimenta tekstova na srpskom predstavljena su dva anotirana korpusa tekstova iz domena filmova – jedan na nivou dužih dokumenata, i drugi na nivou kratkih komentara. *SerbMR* [16] je izbalansiran korpus filmskih recenzija koji je dostupan u varijantama sa dve klase polarnosti sentimenta (pozitivna i negativna) i sa tri klase polarnosti (pozitivna, neutralna i negativna). Dvoklasna varijanta *SerbMR-2C* sadrži ukupno 1682 dokumenta, a troklasna *SerbMR-3C* ukupno 2523, odnosno 841 dokument po klasi. Oznake sentimenta u ovom skupu podataka su dobijene automatski, konverzijom numeričkih ocena pridruženih svakoj recenziji od strane njenog autora.

SentiComments.SR korpus kratkih tekstova na srpskom jeziku [17] je ručno anotiran oznakama sentimenta koje omogućavaju više nivoa interpretacije. Zbog toga se ovaj korpus može upotrebiti u obučavanju i evaluaciji klasifikatora na većem broju užih problema u analizi sentimenta, uključujući određivanje polarnosti, određivanje subjektivnosti, detekciju sarkazma, itd. *SentiComments.SR* sadrži 3490 kratkih komentara i zajednički je anotiran od strane dvoje anotatora.

C. Veb korpus

Veb korpus *srWaC* [18], [19] predstavlja najveći javno dostupni korpus tekstova opšteg tipa na srpskom jeziku. U aktuelnoj verziji 1.1, on sadrži 555 miliona tokena i preko 25 miliona rečenica raspoređenih u oko 1,3 miliona dokumenata. Ovaj korpus je izgrađen prikupljanjem celokupnog sadržaja sa *.rs* domena, nakon čega je sprovedeno uklanjanje duplikata na nivou pasusa, vraćanje dijakritičkih oznaka u tekstovima, kao i automatsko morfosintaktičko označavanje i lematizacija.

III. Alati i računarski modeli

U ovom odeljku su opisani javno dostupni alati i modeli koji se mogu koristiti za rešavanje konkretnih NLP zadataka u obradi tekstova na srpskom jeziku. Oni su izloženi po rastućem nivou obrade, počevši od osnovnih alata za podelu i korekciju tekstova, preko alata za morfosintaktičku i sintaktičku obradu, do alata i modela vezanih za semantičke probleme. Na kraju je opisan *ReLDIanno* veb servis [20] u okviru koga je integrisan veliki broj navedenih alata.

A. Segmentacija i tokenizacija

Uobičajen prvi korak u obradi tekstualnog sadržaja jeste njegova segmentacija i tokenizacija, tj. podela dokumenata na rečenice, i rečenica na elementarne jedinice – tokene – koji predstavljaju reči, brojeve, znakove interpunkcije, itd. Za tokenizaciju srpskog, hrvatskog, slovenačkog, makedonskog i bugarskog jezika može se koristiti ReLDI tokenizator, koji sadrži odvojene modele za standardne i nestandardne tekstove na svim navedenim jezicima. ReLDI tokenizator pruža izlaz u vertikalizovanom, CoNLL-U formatu, koji predstavlja ulazni format podataka za alate za vraćanje dijakritičkih oznaka, morfosintaktičko označavanje i lematizaciju, sintaktičko parsiranje i označavanje imenovanih entiteta.

B. Vraćanje dijakritičkih oznaka

U obradi latiničnih tekstova na srpskom jeziku prikupljenih sa društvenih mreža, veb foruma, komentara posetilaca na različitim sajtovima, itd. često se susreće odsustvo dijakritičkih oznaka, tj. pisanje tekstova engleskom/ASCII latinicom, bez slova č, ć, š, đ, ž i dž. Nedostatak dijakritika može znatno otežati pravilnu automatsku obradu tekstova, jer dovodi do ortografske istovetnosti različitih, često morfosintaktički i/ili semantički veoma udaljenih reči (npr. glagol *sesti* i redni broj *šesti* se svode na isti oblik). Stoga je preporučljivo pre dalje obrade sprovesti vraćanje nedostajućih dijakritičkih oznaka, za šta se može iskoristiti statistički alat za redijakritizaciju [21], koji, pored srpskog, podržava i hrvatski i slovenački jezik.

C. Morfosintaktičko označavanje

Morfosintaktičko označavanje, tj. automatsko dodeljivanje morfosintaktičkog opisa svakom tokenu u rečenici, što uključuje informacije o vrstama reči, predstavlja važan korak u NLP obradi jer su ovi opisi vredan ulazni podatak za mnoge složenije zadatke. Trenutno najbolje javno dostupno rešenje za morfosintaktičko označavanje srpskog jezika je skup alata centra znanja za južnoslovenske jezike CLASSLA [22], koji predstavlja izmenjenu i proširenu verziju *Stanza* biblioteke [23]. Unutar CLASSLA paketa, za srpski jezik su dostupni model za standardni jezik [24] i model prilagođen nestandardnom jeziku [25]. Posebna snaga ovog i svih drugih modela za nestandardne tekstove u okviru CLASSLA paketa je što su izrazito otporni na izostavljanje dijakritičkih znakova u tekstu. Alat generiše morfosintaktičke opise po MULTEXT-East standardu, ali i po sve popularnijem Universal Dependencies standardu,

po kojem se opis sastoji od univerzalne vrste reči i univerzalnih morfosintaktičkih odlika.

D. Morfološka normalizacija

Morfološka normalizacija tekstova omogućava svođenje različitih oblika reči na istu zajedničku osnovu i moguće ju je sprovesti pomoću stemera ili lematizatora. U okviru paketa *SCStemmers* [16], napisanog u programskom jeziku Java, reimplementirana su četiri algoritma stemovanja koja su primenjiva na srpski jezik:

- Optimalni i pohlepni stemer Kešelja i Šipke [26]
- Unapređenje njihovog pohlepnog stemera koje je predstavio Milošević [27]
- Stemer za hrvatski jezik Ljubešića i Pandžića, koji predstavlja unapređenje pristupa iz [28]

Iako do sada nije sprovedena uporedna intrinzička evaluacija ovih algoritama, ekstrinzičke evaluacije na semantičkim problemima određivanja semantičke sličnosti kratkih tekstova i analize sentimenta [13], [17], [29], [30] konzistentno pokazuju da je stemer Ljubešića i Pandžića obično najbolji izbor.

Trenutno najtačnije otvoreno rešenje za lematizaciju srpskog jezika je lematizator koji je deo CLASSLA lanca alata [22]. On koristi flektivni leksikon *srLex* [31], [32] za sve oblike koji su pokriveni leksikonom. Za ostale oblike alat koristi seq2seq model koji na temelju oblika i njegovog morfosintaktičkog opisa oblikuje lemu. Kao i za morfosintaktičko označavanje, i za lematizaciju su dostupni modeli za standardne [33] i za nestandardne tekstove [34].

E. Sintaktičko parsiranje

Sintaktičko parsiranje predstavlja automatsku izradu sintaksnog stabla rečenice kroz obeležavanje sintaktičkih veza između tokena. Iako se parsiranje ređe koristi kao pretprocesiranje za zadatke višeg nivoa, ono je veoma bitno za potrebe lingvističke analize teksta, a može se upotrebiti i za analizu i interpretaciju rada računarskih modela. Od 2017. godine, srpski jezik je kroz anotaciju korpusa *SETimes.SR* [35] prisutan u međunarodnom projektu standardizacije označavanja sintaktičkih dependencija Universal Dependencies [36], pri čemu se anotirani resursi u UD repozitorijumima regularno ažuriraju. Već pomenuti CLASSLA paket takođe pruža mogućnost i za sintaktičko parsiranje tekstova na srpskom. Za razliku od morfosintaktičkog označavanja i lematizacije, u parsiranju se koristi isti model za obradu i standardnih i nestandardnih tekstova [37].

F. Označavanje imenovanih entiteta

Kao i u slučaju morfosintaktičkog i sintaktičkog označavanja i lematizacije, CLASSLA lanac alata [22] postiže trenutno najbolje rezultate i na problemu označavanja imenovanih entiteta. Pri tome, alat razlikuje istih pet tipova entiteta označenih u *SETimes.SR* i *ReLDI-NormTagNER-sr* korpusima, koji su i korišćeni pri obučavanju modela, i koristi isti IOB2 format oznaka. Isto kao i kod rešenja za morfosintaktičko označavanje i lematizaciju, alat omogućuje obradu standardnih [38] i nestandardnih tekstova [39].

G. Određivanje semantičke sličnosti kratkih tekstova

Za dobijanje granuliranih ocena semantičke sličnosti kratkih tekstova dostupan je paket *STSFineGrain* [13] u okviru koga je implementirano više modela za rešavanje ovog problema, koji se mogu obučiti i evaluirati korišćenjem *STS.news.sr* ili nekog sličnog korpusa na drugom jeziku. To uključuje kako osnovne modele zasnovane na leksičkom preklapanju tekstova i/ili sličnosti vektora značenja reči u njima, tako i nešto naprednije modele koji se oslanjaju i na informacije o frekventnosti reči [15], vrsti reči [40] ili na oba tipa informacija [13].

H. ReLDIanno veb servis

ReLDIanno veb servis [20] omogućava jednostavno korišćenje jezičkih alata za srpski, hrvatski i slovenački jezik, uključujući alate za tokenizaciju, morfosintaktičku analizu, lematizaciju, sintaktičko parsiranje i označavanje imenovanih entiteta. Servisima se može pristupiti putem veb aplikacije ili programski, kroz biblioteku za programski jezik Python. Unutar veb aplikacije omogućena je obrada različitih formata fajlova, kao što su TXT, DOCX, PDF, i ZIP arhive. Alati koji se trenutno nalaze u pozadini servisa predstavljaju, uz izuzetak tokenizatora, stariju generaciju otvorenih alata za obradu srpskog jezika koji su razvijeni unutar projekta ReLDI [1].

IV. Zaključak

U ovom radu su ukratko izloženi aktuelni otvoreni, slobodno dostupni anotirani korpusi, kao i alati i modeli za računarsku obradu tekstova na srpskom jeziku. U planu je dalje proširivanje skupa otvorenih resursa novim vrstama anotiranih podataka, poput anotacija koreferentnih odnosa. Pored toga, u skoroj budućnosti se planira i zamena modela korišćenih u okviru veb servisa *ReLDIanno* novijim verzijama iz CLASSLA paketa [22]. Na kraju, u toku su i prvi eksperimenti sa velikim jezičkim modelima poput BERT-a [41], koji će verovatno predstavljati sledeći korak u razvoju računarske obrade srpskog jezika.

Zahvalnica

Autori su zahvalni velikom broju saradnika i anotatora bez čije pomoći izrada navedenih otvorenih resursa ne bi bila moguća.

Rad na ovoj publikaciji delimično je podržan od strane Ministarstva prosvete, nauke i tehnološkog razvoja, i sproveden uz podršku Fonda za nauku Republike Srbije, projekat 6526093, VI-AVANTES.

Dostupnost podataka: Otvoreni resursi su dostupni preko sajta ReLDI centra za jezičke podatke (https://reldi.spur.uzh.ch/hr-sr/resursi-i-alati) i CLARIN.SI repozitorijuma (https://www.clarin.si). Veb servis *ReLDIanno* (http://www.clarin.si/services/web) je dostupan preko sajta CLASSLA centra (https://www.clarin.si/info/k-centre), dok je novi CLASSLA paket alata dostupan preko *GitHub*-a (https://github.com/clarinsi/classla-stanfordnlp) i *pypi* usluge (https://pypi.org/project/classla).

Literatura

[1] T. Samardžić, N. Ljubešić, and M. Miličević, "Regional Linguistic Data Initiative (ReLDI)," in *Proceedings of the Fifth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, 2015, pp. 40–42

[2] V. Batanović, N. Ljubešić, T. Samardžić, and T. Erjavec, "Training corpus SETimes.SR 1.0." Slovenian language resource repository CLARIN.SI, 2018, [Online]. Available: http://hdl.handle.net/11356/1200

[3] V. Batanović, N. Ljubešić, and T. Samardžić, "SETimes.SR – A Reference Training Corpus of Serbian," in *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, 2018, pp. 11–17

[4] F. M. Tyers and M. S. Alperen, "South-East European Times : A parallel corpus of Balkan languages," in *Proceedings of the Workshop on Exploitation of Multilingual Resources and Tools for Central and (South) Eastern European Languages, LREC 2010*, 2010, pp. 49–53

[5] Ž. Agić and N. Ljubešić, "The SETIMES.HR Linguistically Annotated Corpus of Croatian," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, pp. 1724–1727

[6] N. Ljubešić, Ž. Agić, F. Klubička, V. Batanović, and T. Erjavec, "hr500k – A Reference Training Corpus of Croatian," in *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, 2018, pp. 154–161

[7] N. Ljubešić, Ž. Agić, F. Klubička, V. Batanović, and T. Erjavec, "Training corpus hr500k 1.0." Slovenian language resource repository CLARIN.SI, 2018, [Online]. Available: http://hdl.handle.net/11356/1183

[8] N. Ljubešić and M. Miličević, "Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets," *Slov. 2.0 Empirical, Appl. Interdiscip. Res.*, vol. 4, no. 2, pp. 156–188, 2016, doi: 10.4312/slo2.0.2016.2.156-188

[9] N. Ljubešić, T. Erjavec, V. Batanović, M. Miličević, and T. Samardžić, "Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1." Slovenian language resource repository CLARIN.SI, 2019, [Online]. Available: http://hdl.handle.net/11356/1240

[10] N. Ljubešić, T. Erjavec, V. Batanović, M. Miličević, and T. Samardžić, "Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1." Slovenian language resource repository CLARIN.SI, 2019, [Online]. Available: http://hdl.handle.net/11356/1241

[11] D. Fišer, N. Ljubešić, and T. Erjavec, "The Janes project: language resources and tools for Slovene user generated content," *Lang. Resour. Eval.*, vol. 54, no. 1, pp. 223–246, 2020, doi: 10.1007/s10579-018-9425-z

[12] T. Erjavec *et al.*, "CMC training corpus Janes-Tag 2.1." Slovenian language resource repository CLARIN.SI, 2019, [Online]. Available: http://hdl.handle.net/11356/1238

[13] V. Batanović, M. Cvetanović, and B. Nikolić, "Fine-grained Semantic Textual Similarity for Serbian," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1370–1378

[14] V. Batanović, B. Furlan, and B. Nikolić, "Softverski sistem za određivanje semantičke sličnosti kratkih tekstova na srpskom jeziku," *Zbornik radova sa 19. telekomunikacionog foruma (TELFOR 2011)*, Nov. 2011, pp. 1249–1252, doi: 10.1109/TELFOR.2011.6143778

[15] B. Furlan, V. Batanović, and B. Nikolić, "Semantic similarity of short texts in languages with a deficient natural language processing support," *Decis. Support Syst.*, vol. 55, no. 3, pp. 710–719, Feb. 2013, doi: 10.1016/j.dss.2013.02.002

[16] V. Batanović, B. Nikolić, and M. Milosavljević, "Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 2688–2696

[17] V. Batanović, "Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima," Univerzitet u Beogradu, 2020

[18] N. Ljubešić and F. Klubička, "{bs,hr,sr}WaC –Web corpora of Bosnian, Croatian and Serbian," in *Proceedings of the Ninth Web as Corpus Workshop (WaC-9)*, 2014, pp. 29–35

[19] N. Ljubešić and F. Klubička, "Serbian web corpus srWaC 1.1." Slovenian language resource repository CLARIN.SI, 2016, [Online]. Available: http://hdl.handle.net/11356/1063

[20] N. Ljubešić *et al.*, "Easily Accessible Language Technologies for Slovene , Croatian and Serbian," in *Proceedings of the Conference on Language Technologies & Digital Humanities*, 2016, pp. 120–124

[21] N. Ljubešić, T. Erjavec, and D. Fišer, "Corpus-Based Diacritic Restoration for South Slavic Languages," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 3612–3616

[22] N. Ljubešić and K. Dobrovoljc, "What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian," in *Proceedings of the Seventh Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, 2019, pp. 29–34

[23] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)*, 2020, pp. 101–108, doi: 10.18653/v1/2020.acl-demos.14

[24] N. Ljubešić, "The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Serbian 1.1." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1349

[25] N. Ljubešić and V. Štefanec, "The CLASSLA-StanfordNLP model for morphosyntactic annotation of non-standard Serbian 1.0." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1332

[26] V. Kešelj and D. Šipka, "Pristup izgradnji stemera i lematizora za jezike s bogatom fleksijom i oskudnim resursima zasnovan na obuhvatanju sufiksa," *INFOTEKA - časopis za Bibl. i Inform.*, vol. 9, no. 1–2, pp. 21–31, 2008

[27] N. Milošević, "Stemmer for Serbian language." arXiv 1209.4471, 2012

[28] N. Ljubešić, D. Boras, and O. Kubelka, "Retrieving Information in Croatian: Building a Simple and Efficient Rule-Based Stemmer," in *INFuture2007: Digital Information and Heritage*, Zagreb, Croatia: Department for Information Sciences, Faculty of Humanities and Social Sciences, 2007, pp. 313–320

[29] V. Batanović and B. Nikolić, "Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization and Word Embeddings," *Telfor J.*, vol. 9, no. 2, pp. 104–109, 2017, doi: 10.5937/telfor1702104B

[30] V. Batanović and B. Nikolić, "Sentiment Classification of Documents in Serbian: The Effects of Morphological Normalization," in *Proceedings of the 24th Telecommunications Forum (TELFOR 2016)*, 2016, pp. 889–892, doi: 10.1109/TELFOR.2016.7818923

[31] N. Ljubešić, F. Klubička, Ž. Agić, and I.-P. Jazbec, "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian," in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, pp. 4264–4270

[32] N. Ljubešić, "Inflectional lexicon srLex 1.3." Slovenian language resource repository CLARIN.SI, 2019, [Online]. Available: http://hdl.handle.net/11356/1233

[33] N. Ljubešić, "The CLASSLA-StanfordNLP model for lemmatisation of standard Serbian 1.2." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1355

[34] N. Ljubešić, "The CLASSLA-StanfordNLP model for lemmatisation of non-standard Serbian 1.1." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1351

[35] T. Samardžić, M. Starović, Ž. Agić, and N. Ljubešić, "Universal Dependencies for Serbian in Comparison with Croatian and Other Slavic Languages," in *Proceedings of the Sixth Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017)*, 2017, pp. 39–44

[36] D. Zeman *et al.*, "CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies," in *Proceedings of the*

CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2018, pp. 1–21, doi: 10.18653/v1/K18-2001

[37] N. Ljubešić, "The CLASSLA-StanfordNLP model for UD dependency parsing of standard Serbian." Slovenian language resource repository CLARIN.SI, 2019, [Online]. Available: http://hdl.handle.net/11356/1260

[38] N. Ljubešić, "The CLASSLA-StanfordNLP model for named entity recognition of standard Serbian 1.0." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1323

[39] N. Ljubešić, "The CLASSLA-StanfordNLP model for named entity recognition of non-standard Serbian 1.0." Slovenian language resource repository CLARIN.SI, 2020, [Online]. Available: http://hdl.handle.net/11356/1341

[40] V. Batanović and D. Bojić, "Using Part-of-Speech Tags as Deep-Syntax Indicators in Determining Short-Text Semantic Similarity," *Comput. Sci. Inf. Syst.*, vol. 12, no. 1, pp. 1–31, 2015, doi: 10.2298/CSIS131127082B

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423