Duško Vitas

O RAČUNARSKOJ OBRADI SRPSKOG JEZIKA

1. Uvod

Prvi impuls za računarsku obradu srpskog (u ono doba srpskohrvatskog) jezika došao je sa skupa „Kompjuterska obrada lingvističkih podataka“ koji je organizovao u Sarajevu 1977. godine Milan Šipka [17]. Ova konferencija je pokazala da se u to doba u Sloveniji već radilo na složenim obradama slovenačkog jezika, da je u Hrvatskoj bilo različitih aktivnosti oko formiranja elektronskih korpusa, ali da su istraživanja u Srbiji bila ograničena na onovremene metode prepoznavanja govora, tj. prepoznavanja izolovanih fonema.

Pod uticajem ove konferencije formljen je 1978. godine u Matematičkom institutu SANU „Seminar za matematičku i računarsku lingvistiku“ kao mesto okupljanja informatičara i lingvista. Ovaj seminar, koji živi u različitim oblicima i danas[1], stvarni je početak okupljanja istraživača različitih profila, sa ciljem da se razmotre mogućnosti i pravci obrade našeg jezika. Već u prvim godinama su – pored domaćih učesnika, uključujući tu i one sa prostora bivše Jugoslavije – gostovali predavači iz Nemačke, Češke, Rusije, Francuske i drugih zemalja. Ovi prvi koraci[2] ka automatskoj obradi teksta na našem jeziku bili su zabeleženi u časopisu *Prevodilac* [19].

Na inicijativu Evropskog saveta, formiran je 1989. godine kooperativni projekat *Jezičke industrije,* sa ciljem da podstakne saradnju između zapadnoevropskih i istočnoevropskih istraživača sa ovog područja. Na prvom susretu u Dubrovniku učestvovala su velika imena lingvistike sa obe strane gvozdene zavese: Džon Sinkler[3], Moris Gros, Antonio Zampoli, Volfagng Tojbert, Ferenc Kifer, Januš Bienj, Eva Hajičova, a ovaj i potonji susreti su omogućili i domaćim istraživačima da se uključe u najvažnije evropske tokove u ovoj oblasti. Iz ove inicijative je nastao evropski projekat *Telri*[4] čiji je jedan od osnovnih ciljeva bio da se oforme standardi za opis jezičkih resursa za jezike Istočne i Centralne Evrope, a potom i da se ovi standardi primene u njihovoj izgradnji. Za nas je to bilo vreme sankcija gde su, pored svega ostalog, stradala i istraživanja,

1 http://jerteh.rs/?page_id=738

2 U to doba još nije bilo ni ličnih računara, ni računarskih mreža.

3 Biografije nekih od učesnika ovih projekata se i danas nalaze na http://telri.nytud.hu/telri-partners.html

4 http://telri.nytud.hu/

a u ovim aktivnostima smo mogli učestvovati uz izvesna ograničenja. Tek od 2000. godine obnavlja se življa aktivnost na ovom polju, i to prvo u okviru projekata „Interakcija gramatike i rečnika", na kome su učestvovale Katedre za srpski jezik iz Beograda i Novog Sada i informatičari sa Matematičkog fakulteta u Beogradu. Najvažniji rezultat ovog projekta je, verovatno, bilo formiranje prvog elektronskog korpusa savremenog srpskog jezika koji je od tada dostupan preko veba. Za ovaj projekat je vezan i početak razvoja drugih resursa i alata za obradu srpskog jezika. Zatim su sledili brojni domaći i međunarodni projekti čije ćemo osnovne rezultate predstavati u ovom tekstu.

Jezičke industrije obuhvataju danas široko područje u kome se grade programski sistemu namenjeni nekom vidu komunikacije prirodnim jezikom sa računarom. Bilo da je reč o jednostavnim programima – npr. za otkrivanje grešaka prilikom unosa teksta pomoću računara ili o vrlo složenim sistemima kao što su programi za automatsko prevođenje ili glasovnu komunikaciju sa mašinama – neophodno je raspolagati formalizovanim opisom jezika koji se obrađuju. Ovakav opis ima više komponenata – rečnike različitih vrsta, raznovrsne korpuse koji predstavljaju ponašanje jezika u određenom domenu kao i gramatike sa formalizovanim opisom sintaksičkih struktura jezika. Uz ove komponente treba raspolagati i različitim softverskim sistemima koji omogućavaju izgradnju i eksploataciju ovakvih resursa. U ovom tekstu ćemo opisati osnovne resurse i na njima zasnovane aplikacije koji su tokom poslednjih decenija razvijeni za srpski jezik u okviru Grupe za jezičke tehnologije, koja danas radi pod okriljem Društva za jezičke resurse i tehnologije[5].

2. Osnovni zadaci u obradi jezika

Metode automatska obrade jednog teksta (ili korpusa) pomoću računara nastoje da amorfni niz karaktera opreme jezičkim informacijama. Iako korisnik „vidi" na ekranu poruku koju namerava da unese, interno, tekst koji se unosi u računar (u računarskoj memoriji) nije ništa drugo do jedna sekvencija karaktera bez ikakve jezičke organizacije. Jedan jednostavan primer pokazuje ovu razliku između onoga što korisnik vidi i onoga što se nalazi u računarskoj memoriji: jedno slovo, npr. veliko A, koje ima isti oblik u različitim sistemima pisanja (ćirilica, latinica, grčki alfabet) na nivou interne reprezentacije u računaru predstavlja različite objekte. Korisnik koji vidi na ekranu to slovo A, interpretira ovaj grafički oblik u skladu sa jezikom poruke koju piše ili čita, dok računar, budući da „ne zna" ništa o jeziku poruke, dopušta da se izmešaju prilikom pisanja u istoj reči elementi različitih sistema pisanja. U ovom smislu uneti tekst je amorfni niz karaktera bez jezičke interpretacije. U jednoj davnašnjoj definiciji teksta, koju je dala Međunarodna organizacija za standarde (ISO)[6], tekst je „informacija koja se može prikazati u dvodimenzionalnom

5 http://jerteh.rs/

6 ISO/IEC JTC1/SC18 no. 292, p. 6

obliku na ekranu računara ili na papiru, namenjen ljudskom sporazumevanju, a koji se sastoji od karaktera" itd. Drugim rečima, računar u koji se unosi tekst se javlja kao komunikacioni kanal na isti način na koji je to, na primer, vazduh u prenošenju poruke glasom. Opremanje tog niza karaktera (za koji se pretpostavlja da je organizovan jezikom, ali se o njegovoj jezičkoj organizaciji ne zna ništa unapred) jezičkim informacijama je osnovni predmet oblasti koja se naziva (računarska) obrada jezika.

Ovaj niz karaktera se može opremati jezičkim informacijama na različitim nivoima [5]. Prvi korak je da se u njemu identifikuju potencijalni jezički objekti – potencijalne reči jezika, brojevi, interpunkcijski znaci (oni se zovu *tokeni*), a sam proces njihovog izdvajanja se označava kao *tokenizacija* teksta. Iako ovaj korak izgleda jednostavan, on zavisi veoma od sistema pisanja, ali i od različitih konvencija koje se koriste u zapisivanju teksta. Ako potencijalne reči definišemo kao nizove slova određenog sistema pisanja, onda će pojavljivanje nekog niza latiničnih slova u ćiriličnom tekstu biti token, ali ne nužno i reč dotičnog jezika. Složeniji slučaj predstavlja „reč" *3%-ni rastvor* koja se sastoji od tokena *3* koji je cifra, simbola *%* i crtice, i fiktivnog pridevskog nastavka *ni*. Ovakve „nereči", u stvari nizovi tokena različite prirode, postoje u većini tekstova, ali njihovu strukturu i jezički status nije moguće unapred predvideti i opisati.

Sledeći korak u opremanju tokena jezičkim informacijama sastoji se u tome da se potencijalnim rečima (kao nizovima alfabetskih simbola) dodeli uobičajeni rečnički oblik – oblik rečničke odrednice ili *lema*. Tako, na primer, tokenu *crtam* treba dodeliti infinitiv *crtati*, dok bi fiktivno *ni* iz gornjeg primer bilo označeno, u najboljem slučaju, kao nepoznata reč čija se lema ne može odrediti. Ovaj proces se naziva *lematizacija*. Prateći korak uz lematizaciju je i pokušaj da se rečima iz teksta automatski dodeli informacija koja opisuje njen morfosintaksički odnos sa lemom. Na primer, uz oblik *crtam* treba dodati informaciju da je to oblik prvog lica jednine prezenta glagola *crtati*. Ovakva informacija se naziva obično *obeležje* ili (prema engleskom) *tag*, a sam proces nazivamo *obeležavanje*/ *tagiranje* teksta.

Ova tri osnovna koraka, kada se primene na tekst, treba da proizvedu za svaku reč tri oblika (up. *crtam, crtati, 1. lice jednine prezenta*). Ali, ovako zabeležena trojka je neprikladna za dalje informatičke obrade, pa su zbog toga razvijeni različiti standardi za njeno formalizovano opisivanje. Jedan rani standard (razvijen u okviru projekta *Multext-East*) kodira ovu informaciju šifrom *Vmr1s* gde je *V* oznaka za glagol, *m* označava da je reč o leksičkom, a ne pomoćnom glagolu, a kod *r1s* je zamena za neformalizovani opis tj. za *1. lice jednine prezenta*. Kôd za nominativ singulara imenice muškog roda glasi u istom sistemu kodiranja *Npmsn*, pa vidimo da isto slovo (m) kodira različite informacije u zavisnosti od toga da li je reč o glagolu ili o imenici. U jednom drugom sistemu kodiranja – koji je korišćen u opisivanju ovakvih informacija u srpskom jeziku, a o kome će kasnije biti reči – za svaki element morfosintaksičke informacije je izabrano drukčije slovo. Način izbora ovakvih kodova

određuje ne samo uobičajena gramatičko obeležavanje, već i priroda programskih alata koji će se koristiti za obradu teksta.

S druge strane, odnos između oblika reči iz teksta i njenog kanonskog, rečničkog oblika nije uvek precizno definisan, ma kako to izgledalo neobično. Ako pogledamo u tri velika rečnika srpskog jezika (up. RSANU, RMS i JRMS)[7], vidimo da se gramatička informacija u svakom od njih oslanja u velikoj meri na korisničko poznavanje jezika, što znači da odrednica nije uvek standardizovana. Tako (u RSANU) odrednica glasi *vodozemac*, a u RMS – *vodozemci*. Za razliku od druga dva rečnika, gde je kanonski oblik prideva u neodređenom vidu muškog roda, a glagolska imenica posebna lema, u (JRMS) je za odrednicu najčešće biran određeni vid prideva muškog roda, a glagolska imenica je podvedena pod glagol. Kada je u pitanju gramatička informacija, onda podatak o animatnosti nije nikada naveden, premda on određuje oblik akuzativa kod nekih imenica i može biti presudan u razdvajanju značenja. Pored toga, prirodni rod i broj nisu obeleženi. Na primer, imenica *dvojica* je u (RSANU) obeležena kao imenica muškog, u RMS kao ženskog roda, a u *Pravopisnom rečniku* stoji samo uputstvo da se menja kao *stolica* u jednini. Neophodne informacije o kongruenciji koje bi omogućile gramatičko tumačenje niza *ta dvojica su došla*... nema ni u jednom domaćem rečniku.

Ni semantički markeri se ne dodeljuju sistematično. Tako je, na primer, u 20 tomova (RSANU) svega četrdesetak imenica obeleženo oznakom *kul*(inarstvo), pa uz odrednicu *pasulj* stoji da je reč o botaničkom i agronomskom pojmu, ali ne i kulinarskom, premda se većina primera odnosi na *kuvani pasulj*. Oblici koji odstupaju od nemarkirane promene se navode iscrpno iza odrednice. Ali, i tu nedostaju kriterijumi koji bi precizno specifikovali flektivnu paradigmu. Tako se za imenicu *otac* navode ravnopravno množine *očevi* i *oci* (pored *ocevi*), premda se značenja parova *otac, očevi* i *otac, oci* jasno razlikuju. Ovih nekoliko primera pokazuje da konverzija postojećih rečnika u oblik koji bi omogućio automatsku lematizaciju i tagiranje nije najsrećnije rešenje. Pod ovakvim uslovima, proces formalizacije osnovnih koraka u obradi teksta na srpskom mora započeti uz temeljno preispitivanje činjenica koji nude tradicionalni rečnici (i gramatike) [8].

Pored određivanja leme i obeležja, već u prvim koracima treba odrediti i potencijalne granice rečenica u tekstu, što je svojevrsni paradoks u obradi jezika. Utvrđivanje činjenice da li neki niz karaktera predstavlja rečenicu, pre nego što se ispita njegova sintaksička struktura, naoko je totalno besmislen pokušaj. Međurim, da bi se analiza ulaznog teksta mogla ograničiti na nivo rečenice, to se različitim dosetkama određuju (manje-više uspešno) potencijalne granice rečenice. Svaki od navedenih koraka podrazumeva da će odgovarajuće znanje o jeziku biti predstavljeno u okviru nekog formalnog (matematičkog) modela.

7 Reference na ove rečnike su date na kraju teksta.

Iza ovih prvih koraka obrade dolaze složeniji zadaci koji zavise od cilja same obrade. Tekst se može analizirati da bi bio apstrahovan: sveden na kratku poruku o onome što je njegov sadržaj, preveden na neki drugi jezik ili klasifikovan prema nekom kriterijumu koji nije eksplicitno naveden u samom tekstu, itd. Za ovaj poslednji problem, jedan primer pruža zadatak klasifikacije tekstova prikupljenih sa veba za koje treba automatski utvrditi da li predstavljaju, na primer, govor mržnje ili lažne informacije [15]. Među zadacima klasifikacije je i određivanje jezika na kome je tekst napisan. Ovo je posebno zanimljiv problem kada se radi o bliskim standardnim jezicima kao što je to slučaj sa jezicima koji su se razvili iz nekadašnjeg srpskohrvatskog jezika [24].

Ovakve analize treba da omoguće i pouzdanu pretragu velikih zbirki dokumenata. Na primer, kada se zadaje upit nekoj pretraživačkoj mašini kakva je *gugl*, onda očekujemo da nam mašina ponudi relevantna dokumenta koja sadrže tražene reči. Međutim, ako se pretraga obavi doslovno, može se desiti da ne dobijemo nijedan odgovor, premda smo sigurni da bi moralo negde biti dokumenata koje su nam potrebni. Jedan primer koji ovo ilustruje je stigao iz analize upita koje su korisnici postavljali jednoj beogradskoj pretraživačkoj mašini pre desetak godina. Među upitima se zateklo i pitanje „plate u republikama bivše Jugoslavije“ za koje je svakom govorniku srpskog jezika sasvim jasno kakav se odgovor očekuje. Ipak, u velikoj kolekciji potencijalno relevantnih dokumenata nije pronađen nijedan koji odgovara doslovno ovom nizu reči. Šta više, doskora ni *gugl* nije upućivao ni na jedan takav dokument (osim naših radova u kojima je analiziran ovaj problem). Naime, ako tražite tekst u kome se pojavio niz karaktera (npr. *„plate u republikama bivše Jugosavije*“), onda takvog dokumenta nema među onima kojima *gugl* raspolaže. Ipak, izvesno je da bi ih moralo biti, ali sa drukčijom formulacijom. Možda je *plata* u nekom drugom padeškom obliku, npr. *platama*, možda ima dokumenata koji, umesto o *platama,* govore o *zaradama*, *primanjima*, *ličnim dohocima*, a umesto *bivše Jugoslavije* se možda pominje *ex-Yu* ili *teritoriju bivše Jugoslavije*, itd. Preformulisanje polaznog upita u seriju upita koji će obuhvati moguće morfološke ili leksičke varijacije daće izvesno bolji rezultat – bar neki odgovor na postavljeni upit. I doista, *gugl* će za drukčiji upit “*plate u ex-yu*“ pronaći stotinak relevantnih dokumenata. Ovakvo proširenje polaznog upita je još jedan zadatak koji se može rešiti metodama koje se koriste u obradi jezika.

Osim ovakvih analiza, tekstovi su predmet i filoloških istraživanja kako lingvističkih, tako i u raznim drugim domenima od kojih je možda najzanimljivija primena informatičkih metoda u jednojezičnoj ili višejezičnoj leksikografiji. Pitanje formalizovanja jednog prirodnog jezika se danas rešava na jedan od dva osnovna načina: (1) metodama koje su zasnovane na pravilima ili (2) metodama koje koriste statističke podatke o jeziku. Između ova dva osnovna prilaza postoje i različiti hibridni prilazi koji kombinuju pojedine elemente iz ova dva osnovna prilaza. (Prilaz zasnovan na pravilima će biti ilustrovan kasnije u ovom tekstu.) U grupi statistički zasnovanih prilaza su različiti postupci zasnovani na metodama mašinskog učenja ili na neuronskim

mrežama, a cilj im je da odgovore, pre svega, na rastuće zahteve jezičke industrije podstaknute zahtevima da se brzo analiziraju velike količine teksta sa veba. U osnovi, ovakve metode su neka vrsta crne kutije: jezička opravdanost postupka obično ne postoji, ali se ipak dobijaju korisni rezultati, brzo i po niskoj ceni. Na ovaj način su dobijeni brojni rezultati čija je praktična vrednost nesporna i nivo tačnosti najčešće vrlo visok, često preko 95%. Problem sa jezičkog stanovišta potiče otuda što onaj mali procenat netačnih rezultata prikriva odsustvo znanja o jeziku koji se obrađuje. U popisu literature [10] prikazan je eksperiment sa evaluacijom rezultata različitih programa za morfološko tagiranje teksta na srpskom. Evaluacija je izvršena tako što su programi „učili" na 9/10 jednog ručno tagiranog korpusa šta je imenica, a šta glagol, itd, a onda je rezultat učenja primenjen na poslednju desetinu korpusa. I dok je opšta uspešnost ovih programa bila zadovoljavajuća, dotle je efikasnost na onim rečima kojih nije bilo u skupu na osnovu kojih je tager obučavan drastično pala na skromnih 40-50% tačnih pogodaka. Ovakav rezultat se lako može razumeti ako se ima u vidu popunjenost teksta visokofrekventnim slojevima reči. Na primer, u korpusu nedeljnika „Vreme", koji sadrži oko 30 miliona tokena, jednu petinu svih reči čini svega 10 oblika, a 100 najčešćih oblika reči trećinu svih oblika u korpusu. Jasno je tada da će se tih 100 najfrekventnih reči naći u svakom delu korpusa i da će one biti ispravno obeležene, a ono što nije frekventno će često biti pogrešno obeleženo, ali to ne smanjuje u velikoj meri preciznost programa koji vrši automatsku lematizaciju ili morfološko obeležavanje. Kao primer može da posluži jedan korpus srpskog u kome su leme i morfološka obeležja konstruisane statističkim metodama, gde će se pojaviti konstruisan glagol čiji je infinitiv *vati* sa oblicima paradigme *vamo*, *vao*, *vajući*... koje, verovatno, predstavljaju pogrešno zapisane reči u korpusu. Naravno, paradigmi glagola *vati* nedostaju oblici njegovog prezenta *vam* i *vaš* jer su se oni našli (kao zamenice) u visokofrekventnom sloju na osnovu koga je program „učio" o morfologiji srpskog jezika. Kada se koriste rezultati statističke obrade, treba voditi računa o pojavi neadekvatnih rezultata u niskofrekventnim slojevima korpusa.

3. Račun sa rečima

U okviru formalne teorije jezika [21] opisana je jedna klasa jezika koja omogućava jednostavan račun sa rečima. Reč je o takozvanim *regularnim* ili *racionalnim jezicima*. Ovaj drugi naziv potiče otuda što se reči ovih jezika podvrgavaju većini zakona koji važe i za racionalne brojeve ili, jednostavnije, za razlomke. U najjednostavnijim crtama, to izgleda ovako: zamislimo da imamo azbuku sa samo dva slova, a i b.[8] Reči koje se mogu formirati pomoću ta dva slova su

a, b, aa, ab, ba, bb, aaa, aab, aba, abb...

8 Ne treba napominjati da je takva azbuka dovoljna da se opišu svi mogući sadržaji računarskih memorija, dakle i svih tekstova na bilo kom jeziku koji su napisani uz pomoć računara.

Regularni jezici su neki podskupovi ovog skupa, dakle reči koji se konstruišu pomoću tri jednostavne operacije. To su:

- dopisivanje: od dve reči pravi se treća tako što na prvu „dopišemo“ drugu. Na primer, gornja reč *ba* se dobija tako što na reč *b* dopišemo reč *a;*
- sabiranje: ako su *x* i *y* neke dve reči, onda će reč biti i reč koja se dobija njihovim okupljanjem u jednu reč obeleženu sa *x+y*. Od gorepomenutih reči *ab* i *bb* može se napraviti nova reč *ab+bb* (sa značenjem ili je „izrečeno je *ab* ili je *bb*“);
- ponavljanje ili iteracija: svaka reč takvog jezika se može dopisati na samu sebe neograničen broj puta: ako je *x* neka reč ovog jezika, onda su to i *xx, xxx, xxxx*..., a taj skup reči obeležavamo sa x*.

Pored toga, reč ovog jezika je i prazna reč koju obeležavamo sa ε, a koja se može shvatiti kao nulti-nastavak, npr, u deklinaciji. Ova krajnje pojednostavljena, ali i dalje „suva“ matematika kojom se opisuju regularni jezici ima svoju živu lingvističku interpretaciju, o čemu svedoče sledeći primeri.

Imenička deklinacija se obično opisuje tablicama u kojima su prototipske imenice izmenjane po padežima. Kao primer za imenice na konsonant muškog roda obično se uzima imenica *jelen*. Umesto u obliku tablice, deklinirani oblici se mogu opisati i regularnim izrazom:

jelen + jelena + jelenu + jelena + jelene + jelenu + jelenom +...

a zatim se svakom obliku može dodati gramatički kod koji opisuje njegovu vezu sa kanonskim oblikom *jelen*, npr. jelena/m2s;m4s znači da je oblik jelena genitiv (2) ili akuzativ (4) singulara (s) imenice muškog roda (m) *jelen*. Sa ovim oznakama gornji izraz postaje

jelen/m1s + jelena/m2s;m4s + jelenu/m3s;m7s + jelenom/m6s +...

Kako je reč o regularnom izrazu, za njega važe zakoni distribucije[9], pa se ovaj izraz može zapisati i kao

jelen(ε/m1s +a/m2s;m4s + u/m3s;m7s + om/m6s +...) ...(*)

Ovaj red sadrži iste informacije kao i tabele kojima se obično prikazuje promena imenica. S druge strane, reč *jelen* se ovde može zameniti bilo kojom drugom imenicom nemarkirane promene 1. vrste, dok izraz u zagradama ostaje isti. Kod imenica muškog roda na konsonant kod kojih je osnova alternirana, gornja operacija daće drukčiji izraz. Ovo znači da je na nivou flektivnih reči moguće definisati relaciju „flektivne jednakosti“: dve flektivne reči su „flektivno jednake“ ako se njihova promena predstavlja istim regularnim izrazom oblika analognog onom iz izraza (*). Ako se formiraju takvi izrazi za sve flektivne reči u srpskom, saznaćemo da za imenice postoji oko 400, za glagole

9 Kao što u aritmetici važi da je 3*2 + 3*5 isto što i 3*(2+5).

oko 390 itd. ovakvih različitih regularnih izraza koji precizno opisuju, pored nemarkirane promene, i sve „izuzetke“ našeg jezika [20].

Kao primer za operaciju iteracije posmatrajmo uzvik *ej!* U korpusima srpskog jezika, ovaj uzvik se javlja još u obliku *eej! eeeej! eee...ej!* (čak i do 20 ponovljenih e!) Čitalac će znati da svede ove različite oblike na osnovni oblik u kome je ovaj uzvik prikazan u rečnicima, ali ne i računar. Kako bi i računar to mogao da uradi, potrebno je ovaj uzvik prikazati u obliku regularnog izraza

$e(\varepsilon + e + ee + eee + ...)j$ ili, kraće, $e(e)^*j$.

4. Elektronski rečnici i semantičke mreže srpskog jezika

Za razliku od rečnika (bilo da su na papiru ili u digitalnom obliku), namenjenih ljudskoj upotrebi, elektronski rečnici (kraće, e-rečnik) se konstruišu za obradu jezika na računaru. Koraci koje je potrebno napraviti u početnoj obradi teksta – tokenizacija, tagiranje, lematizacija – mogu se obaviti na osnovu sadržaja e-rečnika metodama leksičkog prepoznavanja [3], [6]. U svom osnovnom obliku e-rečnik ima sledeću strukturu:

w_t, w_l.T:mgp

gde je w_t oblik reči, w_l – odgovarajuća lema, dok je *T* oznaka za vrstu reči, a *mgp* kôd koji opisuje morfološki odnos između w_l i w_t. Na primer, jedan red u rečniku srpskog koji opisuje genitiv ili akuzativ singulara imenice *jelen* imaće oblik.

jelena, jelen.N:m2s:m4s ... (**)

Ukoliko raspolažemo opisom regularnih izraza tipa (*), onda je iz spiska rečničkih odrednica moguće generisati automatski rečnik svih oblika reči u formatu (**). Ovakva konstrukcija se zasniva na tome što se lemi dodeljuje odgovarajući regularni izraz oblika (*) koji određuje njena flektivna svojstva.

Gornji format može biti proširen i na oblik

w_t, w_l.*T + SSP:mgp*,

gde je *SSP* skup sintaksičkih, semantičkih i drugih svojstava leme. Polje *SSP* je proizvoljne dužine i sadrži različite tipove kvalifikatora koji omogućavaju dalje analize teksta. U ovom polju se, na primer, koriste oznaka +MG za prirodni muški rod, pa je imenica *dvojica* opisana kao *dvojica.N+MG+Hum:f1s* gde je *+Hum* marker za *osobu*, a *f* govori da je ova imenica ženskog gramatičkog roda. Još neki markeri koji se koriste su: +VN za glagolske imenice (npr. *generisanje, N+VN*), +Imperf za nesvršeni glagol (npr. *ići,.V+Imperf*) itd. Kod predloga je naveden kôd koji opisuje zahtevani padež (npr, *iza,.PREP+p2* govori da se iza predloga *iza* traži oblik genitiva). Pored ovih morfosintaksičkih kodova, u e-rečniku srpskog su obeležena i vlastita imena: npr. red *Beograd,.N+NProp+Top+Gr+CC2=RS* kazuje da je *Beograd* toponim (*Top*) koji označava grad

(*Gr*) u Srbiji (*CC2=RS*). Tu su i razna druga leksička svojstva (npr. red *crven,. A+Col* kazuje da ovaj pridev označava boju (*Col*)). Aktuelni obim ovog rečnika je oko 200.000 odrednica sa više od šest miliona njihovih morfoloških oblika.

Tagiranje oblika na ovaj način donosi jedan problem kada su u pitanju polileksemske reči. Na primer, sekvenci *s vremena na vreme* će biti pripisane vrste reči *predlog imenica predlog imenica*, ali se tu radi o priloškoj sintagmi. Ovaj problem se rešava konstruisanjem rečnika polileksemskih jedinica čiji je format isti kao (**), ali je dopušteno da se između nizova slova pojave karakteri koji nisu slova (kao što je npr. razmak ili crtica). Ovakvim rečima odgovaraju sledeći redovi rečnika:

s vremena na vreme,.ADV+Comp
s tačke gledišta,.PREP+Comp+p2
naučno-istraživački rad,.N+SIN=2XAXN(sin)+ Comp

gde *Comp* označava polileksemsku jedinicu, a *SIN=2XAXN(sin)* opisuje regularni izraz slaganja prideva sa imenicom u ovoj sintagmi.

Ipak, i sa ovako obimnim rečnikom, u svakom tekstu se javljaju nove reči kojih nema u rečniku [23]. Izvori novih reči su različiti, a jedan od čestih razloga njihovog javljanja je rezultat derivacionih procesa. Jedan broj ovakvih reči – koje imaju status nepoznate reči, sa specifičnim regularnim izrazima koji opisuju derivacione procese – može se prevesti u prepoznate reči, pa tako dobiti odgovarajuću lemu i gramatičku informaciju. Ali, pogledajmo sledeći primer. U korpusu nedeljnika *Vreme* između 2009. i 2018. godine, reč *Tviter* se pojavila prvi put 2009. godine. Tokom ove decenije iz nje se razvilo još 15 novih reči među kojima su *tvit, tviteraš, tviterašica, tviteraški, tviterski, tvitovati, tvitnuti, tvitati*. Ako se *Tviter* i *tvit* unesu kao odrednice u rečnik, tada i izvedenice mogu dobiti odgovarajući status, i to automatski. Ovakva mogućnost u održavanju rečnika je zanimljiva i iz ugla višejezične leksikografije. Naime, kada su u pitanju derivacioni procesi čiji rezultat daje novu odrednicu sa značenjem koje je predvidivo na osnovu polazne reči, onda je moguće flektivnu paradigmu proširiti i derivatima, što vodi ka rečniku ustrojenom prema „super-lemama".

Takva super-lema imenice okuplja iz nje izvedene prisvojne i relacione prideve, deminutive i augmentative, eventualno mociju roda itd, a posledica njenog formiranja je dvostruka: za lemu se vezuje njen ukupan derivacioni potencijal, tj. oblici koji nisu potvrđeni, ali su derivaciono mogući, s jedne strane, dok je takva super-lema, s druge strane, idealni kandidat za lemu u srpskoj koloni dvojezičnog rečnika jer ga ne opterećuje suvišnim odrednicama. Tako se uz imenicu *glumac* u (RMS) navode odrednice *glumački* (kao pridev i prilog), *glumica*, *glumičin*, *glumčev*, *glumčina*, *glumčić*, ali ne i augmentativ i deminutiv za *glumica*, niti pridevi izvedeni iz njih. Rad na srpsko-engleskom rečniku bi, polazeći od spiska odrednica iz (RMS), naišao na ozbiljne probleme koji bi se uvođenjem „super-leme" mogli prevazići [27].

Semantičke relacije su u e-rečniku navedene delimično u okviru polja SSP, ali semantičke veze između lema nije moguće ugraditi u ovom formalizmu, pa su za opis semantičkih relacija konstruisana dva druga resursa. Jedan od njih predstavlja srpsku leksiku u okviru višejezične semantičke mreže *WordNet* [2]. Ova mreža, koja je prvobitno razvijena za engleski jezik, je pomoću više evropskih projekata adaptirana za većinu evropskih jezika, uključujući i srpski. Ova mreža organizuje skupove bliskoznačnih reči u čvorove jednog grafa, u jedno *gnezdo*[10], i uspostavlja veze sa drugim čvorovima preko lukova koji se interpretiraju kao semantičke relacije. Veza između različitih jezika se ostvaruje interlingvalnim indeksom koji numeriše značenjski ekvivalentna gnezda u različitim jezicima. Na primer, za englesko gnezdo čiji je indeks *eng-30-13279262-n*, a koje okuplja bliskoznačne engleske reči *wage, pay, earnings, remuneration, salary,* u srpskom odgovara gnezdo u kome se nalaze *plata, zarada, lični dohodak, nadoknada*. Ovaj čvor je u engleskom jeziku povezan relacijom hiperonimije sa čvorom u kome se nalazi *regular payment*, a u srpskom *redovna isplata*, a derivacionim relacijama još u srpskom sa čvorovima *platiti, isplatiti, zaraditi*. Srpski deo ove mreže sadrži oko 25.000 gnezda sa blikoznačnim rečima[11].

Druga semantička mreža je *Prolex*[12], višejezična mreža vlastitih imena [9]. U ovoj mreži je vlastitim imenima dodeljen identifikacioni broj preko koga se povezuju oblici vlastitih imena u različitim jezicima. Za oblik vlastitog imena u jednom jeziku su vezani sinonimni nazivi, kao i izvedenice (stanovnik, stanovnica, itd). Na primer, za francusku reč *Paris*, srpski ekvivalent sadrži dve lekseme *Pariz* i *Grad svetlosti*, a zatim i nazive stanovnika (*Parižanin*, *Parižanka*, itd). Dalje, svakom vlastitom imenu je dodeljen izvestan broj semantičkih obeležja (npr. imaginarni ili stvarni toponim). Jedna primena ove baze je vidljiva na vebu, na strani projekta *Renom*[13], koji povezuje originalni tekst Rableovog romana *Gargantua* sa kartom stvarnih mesta na koja se Rable pozivao.

5. Programski sistem *Unitex*

Za izgradnju i eksploataciju e-rečnika srpskog jezika se koristi sistem Unitex[14], razvijen na Univerzitetu Marn-la-Vale u Parizu. Početnu verziju sistema je konstruisao Sebastijan Pomije polazeći od ideje leksičkog prepoznavanja i leksikon-gramatika francuskog lingviste Morisa Grosa. Ovaj sistem je dostupan u slobodnoj distribuciji pod licencom LGPLLR, a sadrži module za obradu desetak jezika uključujući srpski (pisan ćirilicom ili latinicom). Ovaj sistem je zasnovan na teoriji čiji je jedan vid navedeni opis jezika posredstvom

10 Na engleskom se ovakvo gnezdo naziva *synset*.

11 Srpski WordNet se može pregledati preko veba na adresi http://dcl.bas.bg/bulnet/, a preuzeti, pod određenim uslovima, na adresi http://korpus.matf.bg.ac.rs/SrpWN/.

12 https://www.ortolang.fr/market/lexicons/prolex/v3.1

13 https://renom.univ-tours.fr/fr

14 https://unitexgramlab.org/

regularnih izraza. Na tekst ili korpus koji se obrađuje, primenjuju se e-rečnici kao i njihove ekstenzije u obliku lokalnih gramatika, a rezultat se dobija u obliku konkordanci ili se polazni tekst transformiše u novi oblik u koji su ugrađene prepoznate informacije. Obrađeni tekst je dalje moguće pretraživati ili transformisati koristeći se bilo kojom kombinacijom elemenata opisanih (uključujući i one u SSP) u e-rečniku. Upit oblika <čovek> izdvojiće iz korpusa sva pojavljivanja u bilo kom padeškom obliku ove leme, a upit <N+MG> sve imenice ženskog roda koje mogu označavati mušku osobu. U pomenutom korpusu nedeljnika *Vreme*, primeri takvih imenica su *sudija, osoba, ubica, dvojica, trojica, gazda, zanatlija, zvanica*... Upit <VN> izdvaja sve glagolske imenice kao *istraživanje, obrazovanje*, *učenje* itd, a upit <A+Col> sve prideve koji se odnose na boje što, pored osnovnih boja kakve su *bela, crna* ili *crvena*, izdvaja i na primere za *svetloplavu, ružičastu, sivomaslinastu, sivkastu, rumenu, azurnu*...

Osim ovih jednostavnih upita na osnovu sadržaja e-rečnika, moguće je upit oformiti na složen način kao lokalnu gramatiku: kombinaciju elemenata iz e-rečnika koja opisuje složeni sintaksički ili semantički obrazac u tekstu [4], [22]. Ovakve gramatike su obično regularni izraz velike složenosti. Jednostavna lokalna gramatika ekstrahuje iz teksta sva pojavljivanja referenci na vreme u bilo kom obliku. Primeri dobijeni ovom gramatikom na korpusu „Vreena“ (a ima ih 32.452) su *od utorka 31. maja uveče*; *u noći između 31. decembra 1958. i 1. januara 1959*;*u ponedeljak, 27. aprila 2015. godine*; *od februara 2004. godine do danas* itd. Jasno je da se sekvence ovakve složenosti ne mogu dobiti prostim pretraživanjem po pojedinačnim rečima. Pomoću lokalnih gramatika moguće je opisati i izdvojiti pojedine rečenične konstituente poput određenih tipova sintagmi, oblika složenih vremena, ispitati raspored enklitika itd, pa je na ovaj način moguće izvršiti i plitku sintaksičku analizu polaznog teksta. Resursi koji su razvijeni u okviru ovog sistema, kao i pojedine njegove funkcije, su integrisani u druge aplikacije koje ćemo pomenuti kasnije.

6. Korpusi srpskog jezika

Pored navedenih leksičkih resursa, konstruisan je i prvi korpus savremenog srpskog jezika koji je istraživačima i studentima srpskog jezika dostupan preko veba. Prva verzija ovog korpusa, postavljena na veb 2002. godine, sadržala je savremene tekstove od oko 23 miliona reči, bez dodatnih obeležja u tekstu, ali sa metapodacima o izvorima. Što se tiče pretrage korpusa, ona se zasnivala isključivo na upotrebi regularnih izraza[15]. Sledeća, znatno proširena verzija ovog korpusa postavljena je na veb 2013. godine [18], [25][16]. Obim ovog korpusa je 122 miliona reči, a proširene su i mogućnosti u pretraživanjatako što je korpus lematiziran, a lemama su dodate i informacije o vrstama reči, kao i podaci o funkcionalnom stilu teksta u kome se pojavljuje pojedina reč. Ovaj korpus ima danas preko 800 korisnika iz zemlje i sveta, a tokom ove

15 http://www.korpus.matf.bg.ac.rs/prezentacija/uputstvo.html
16 http://www.korpus.matf.bg.ac.rs/

godine će biti inoviran, kako u pogledu obima[17], tako i dodatnih informacija koje se mogu koristiti prilikom pretrage.

U nastavku dajemo primer pretrage lematiziranog korpusa pomoću upita [pos="A" & lemma=".*ski"]{3} [pos="N"]. Značenje ovog izraza je sledeće: traži se sekvenca od 3 (*{3}*) prideva (*A*) čija se lema završava na *–ski*, za kojima sledi neka imenica (*N*). U korpusu se nalaze 224 takva primera od kojih su neki *ženski sportski studentski tim*: *srpskog građanskog demokratskog društva; Srpske Kraljevske Dvorske Knjižare, japanski školski ribarski brod, vrhunskog Rimskog carskog mozaika* itd. U odnosu na obradu korpusa pomoću *Unitexa* ovde su mogućnosti značajno ograničene jer se pretrage vrše bez upotrebe e-rečnika, ali se zato mogu pretraživati daleko šire zbirke tekstova.

Pored ovog jednojezičnog korpusa, sastavljeno je više paralelnih korpusa, gde su tekstovi upareni na nivou segmenata (koji aproksimiraju rečenicu). Prvi takav korpus je bio francusko-srpski korpus[18] koji se sastoji od oko 1,5 miliona reči iz svakog od ova dva jezika, a sastavljen je od književnih tekstova iz 19. i 20. veka, kao i od odabranih članaka iz časopisa „*Monde diplomatique*" i dostupan je kroz isti interfejs kao i srpski korpus. Ako u ovom korpusu potražimo francuske ekvivalente za srpsku sekvencu *s vremena na vreme,* dobićemo 87 parova, gde konkordance pokazuju da su česti francuski ekvivalenti *de temps en temps* i *de temps à autre*, ređe *par intervalles* i sasvim retko – *parfois*. Na sličan način je konstruisan i englesko-srpski korpus[19] koji je dostupan preko veba kroz isti interfejs kao i gore opisani korpusi. Njegov obim i struktura je slična onoj francusko-srpskog korpusa. Detaljniji primeri upotrebe ovog korpusa su dati u [7].

Nekoliko domenski definisanih korpusa je u slobodnom pristupu preko sistema *Bibliša*[20]. Tu se nalaze paralelni englesko-srpski korpusi iz različitih specijalizovanih časopisa, kao i nemačko-srpski paralelni korpus koji se sastoji od 14 romana srpskih i nemačkih autora i njihovog prevoda [1]. Svi paralelni korpusi su paralelizovani na nivou rečenice (ekvivalentnih segmenata) i dostupni u formatu *TMX*. Na neke od njih je primenjeno i poravnavanje na nivou reči koristeći sistem *Giza++*. Pored ovih korpusa koji su javno dostupni, izgrađeno je i više malih korpusa za posebne primene. Jedan od njih je mali srpsko-srpski paralelni korpus koji s sastoji od tekstova višestrukih prevoda istog romana na naš jezik. Ovaj korpus pruža primere parafraziranja budući da je isti sadržaj izvornog jezika iskazan, po pravilu, na drugi način. Jedan primer iz ovog korpusa potiče iz višestrukog prevoda romana *Rukopis nađen u Saragosi* Jana Potockog na srpski jezik:

17 Očekuje se da će nova verzija korpusa imati oko 400 miliona reči, a veliki deo korpusa su tekstovi koji se ne nalaze na vebu.

18 http://www.korpus.matf.bg.ac.rs/SrpFranKor/korpus/index1.php

19 http://www.korpus.matf.bg.ac.rs/SrpEngKor/korpus/index1.php

20 http://jerteh.rs/biblisha/

To nije sve. Putnika koji bi se usudio u tu *divlju pokrajinu* *napale bi,* *govorilo se,* hiljade *grozota* *kadrih da slede krv u žilama i najhrabrijih.* *Tako* bi on čuo *žalobne* glasove *pomešane* *sa hučanjem* potoka *i zviždanjem bure,* *zavodile* bi ga lutajuće svetlosti, *a* nevidljive ruke gurale *u ambise bez dna.*	*Sem toga,* putnika koji bi se usudio *da krene* u tu *divljinu* *progonile su* *(kako se pričalo)* hiljade *užasa,* *pred čijim je prizorom drhtala i najhladnokrvnija hrabrost.* On bi čuo *plačne* glasove *koji se mešaju* *sa hukom* potoka, *kroz zavijanje oluje* *mamile* bi ga lutajuće svetlosti, *a* nevidljive ruke *bi ga* gurale *u bezdane ponore*
Prevod s francuskog[21]	Prevod s poljskog[22]

Polazna svrha ovog eksperimentalnog korpusa je bila vezana za traganje za potvrdama regularne derivacije, a zatim i za potrebe formiranja gnezda u mreži WordNet.

Drugi mali korpus je srpsko-hrvatski korpus koji se sastoji od prevoda popularnih književnih dela na ova dva jezika. Građa za ovaj korpus je preuzeta iz ASPAC-korpusa, a svrha paralelizacije je bila da se kvantifikuju razlike koje postoje između dva uzusa [26]. U fazi je izrade dijahroni kulinarski korpus koji se sastoji od etnografske građe, kulinarskih priručnika i drugih izvora koji govore o prehrambenim navikama među Srbima od sredine 19. veka do danas. Ovaj korpus će pružiti uvid u razvoj kulinarske terminologije, ali i kulinarskih običaja. U pripremi je i korpus 100 srpskih romana objavljenih između 1850. i 1920. godine, razvijen u okiru evropske COST-akcije *Udaljeno čitanje*. Tekstovi za ovaj korpus[23] su uglavnom skenirani iz prvih izdanja romana, zatim konvertovani u tekst i poluautomatski korigovani (pomoću *Unitex*a).

21 prevod Slobodana Petkovića, Prosveta, Beograd 1964.
22 prevod Stojana Subotina, SKZ 533-4, Beograd, 1988.
23 https://github.com/COST-ELTeC/ELTeC-srp

7. Aplikacije

Polazeći od opisanih resursa, pre svega leksičkih, razvijene su informatičke aplikacije koje omogućavaju složene eksperimente nad materijalom srpskog jezika. Već pomenuti sistem *Bibliša*[24] [11], [12] omogućava pretragu paralelizovanih kolekcija dokumenata. Osobenost sistema leži u tome što korisnik može upit, postavljen na srpskom, da proširi automatski u gnezdo bliskoznačnih reči iz WordNet-a, dvojezičnih glosara, terminoloških baza itd, da uključi u pretragu njihove hiperonime i hiponime, kao i da izvrši morfološku ekspanziju upita (ne osnovu informacija iz e-rečnika). Za upit *plata* upit se može proširiti rečima *lični dohodak, nadoknada, plaćanje, zarada,* a u pretrazi paralelnog korpusa časopisa *Infoteka* dobijaju se konkordance koje sadrže, pored ostalih, i redove kao u donjoj tabeli:

Mitrović, 2013, vol. XIV:1, ID: 1.2013.1.5 metadata	Another paper (Kazai 2011) describes this type of cooperation as "an open call for contributions from members of the crowd to solve a problem or carry out human intelligence tasks, often in exchange for micro-payments, social recognition, or entertainment value".	Drugi autor (Kazai 2011) ovu vrstu saradnje smatra otvorenim pozivom grupi ljudi da reše neki problem ili da izvrše neki zadatak za koji je neophodna ljudska inteligencija, dakle računari ga ne mogu obaviti, obično u zamenu za malu novčanu **nadoknadu,** sticanje statusa u društvu ili radi zabave.
Kosanović, 2008, vol. IX:1/2, ID: 1.2008.1/2.1 metadata	The aim of the evaluation is to optimize the acquisition and it is achieved by integrating the information on users' needs, service prices, **payment** conditions and degree of utilization of the subscribed resources.	Cilj evaluacije je optimizacija nabavke koja se ostvaruje objedinjavanjem informacija o potrebama korisnika, cenama servisa, uslovima **plaćanja** i iskorišćenosti pretplaćenih izvora.
Zdravkova, 2010, vol. XI:2, ID: 1.2010.2.1 metadata	In smaller markets you don't, you can count on temporary discounts and postponed **payment** with checks (wherever such **payment** method still exists).	U manjim prodavnicama se ne cenkate, ali možete računati na povremena sniženja i odloženo **plaćanje** čekovima (tamo gde takav način **plaćanja** još uvek postoji).
Primer dvojezičnih obeleženih konkordanci u sistemu *Bibliša*		

U ovom primeru rezultat pretrage za ključ *plata* pronalazi u srpskom delu korpusa reči *plaćanje i nadoknada* u različitim padeškim oblicima, kao i engleski ekvivalent *payment*. U prvoj ćeliji tabele *micro-payments* nije obeleženo jer se ne nalazi u engleskom gnezdu u kome je *payment*.

24 http://jerteh.rs/biblisha/

Ipak, najznačajniji segment softverskih alata koji se temelji na razvijenim resursima je namenjen leksikografskom i terminografskom radu. Aplikacija *Termi*[25], koja je nastala prvobitno za potrebe izgradnje terminološke baze za rudarstvo i geologiju, proširena je vremenom na druge domene i predstavlja okruženje za razvoj višejezičnih terminoloških rečnika. Ovu aplikaciju prati sistem *BilTE*[26] za automatsku ekstrakciju kandidata za termine iz dvojezičnih korpusa [16].

Najznačajniji informatički proizvod je sistem *Leximirka*[27] [12], [13], [14] koji je nastao iz konverzije e-rečnika u relacionu bazu podataka, što je omogućilo da se integrišu na jednoj platformi različiti leksički resursi kao što su tradicionalni rečnici u različitim formatima, e-rečnici i korpusi u jedinstveno okruženje namenjeno prvenstveno leksikografskom radu.

8. Zaključak

Opisani resursi za obradu srpskog jezika su nastali prvenstveno iz istraživačke radoznalosti članova Grupe za jezičke tehnologije, a njihovo postojanje načelno obezbeđuje opstanak našeg jezika u digitalizovanom svetu [24]. Ipak, njihova realna primena, nadgradnja i evaluacija zavise u najvećoj meri od njihove buduće upotrebe u realnom korisničkom okruženju. Nažalost, postoje još uvek preveliki otpori različte motivacije, što dovodi do hroničnog zaostajanja za svetskim trendovima. Primer takvog zaostajanja je činjenica da se elektronska knjiga sa srpskim tekstom na *Kindle*-u ili nekoj drugoj e-knjizi ne može koristiti sa onim funkcionalnostima koje ovakvi uređaji pružaju korisnicima na drugim jezicima.

Bibliografija

1. Andonovski, J. *Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa*, doktorska teza, Filološki fakultet, Univerzitet u Beogradu, 2020.
2. Fellbaum C. (ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MIT Press, 1998.
3. Gross, M.; Perrin D. (eds.) *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science 377, Springer-Verlag, 1989.
4. Gross, M. The construction of local grammars. In Roche, E.; Schabes, Y. (eds.), *Finite State Language Processing*, Cambridge, Mass./London: The MIT Press, 1997.
5. Jurafsky, D.; J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall, 2008.

25 https://termi.rgf.bg.ac.rs/
26 http://bilte.jerteh.rs/
27 https://leximirka.jerteh.rs/

6. Krstev, C. *Processing of Serbian : automata, texts and electronic dictionaries*. Belgrade : Faculty of Philology of the University, 2008.
7. Krstev, C.; Vitas D. "An Aligned English-Serbian Corpus, In: *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality),* Volume I, pp. 495-508, Faculty of Philology, University of Belgrade.
8. Krstev, C. O odabiru odrednica za elektronski rečnik srpskog jezika i njihovom povezivanju, *Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene*, Vol. 48/3, Međunarodni slavistički centar, Beograd, 2019, pp. 133-147.
9. Maurel, D., Vitas, D., Krstev, C., Koeva S. Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian. *Bulletin de linguistique appliquée et générale, Presses Universitaires de Franche-Comté*, 2007, 32, pp. 55-72.
10. Popović, Z.. Programi za etiketiranje teksta na srpskom jeziku. INFOteka, 2010, 11(2): 19-36.
11. Stanković R., Krstev, C., Obradović I., Trtovac A., Utvić M. A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2012, pp. 1710-1717.
12. Stanković, R.; Krstev, C.; Lazić, B.; Škorić, M. Electronic Dictionaries—from File System to lemon Based Lexical Database.In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation—W23 6th Workshop onLinked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018*, 2018; pp. 18–23.
13. Stanković, R.; Šandrih, B., Stijović, R., Krstev, C., Vitas, D., Marković, A. SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian.. In *Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference* 2019, Lexical Computing CZ, s.r.o.
14. Stanković, R., Šandrih, B., Krstev, C., Utvić, M., Škorić, M. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, European Language Resources Association.
15. Stanković, R., Mitrović, J., Jokić, D., Krstev, C. Multi-word Expressions for Abusive Speech Detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons* (2020), Association for Computational Linguistics.
16. Šandrih, B.; Krstev, C., Stanković, R. Two approaches to compilation of bilingual multi-word terminology lists from lexical resources. In *Natural Language Engineering* (2020), Cambridge University Press (CUP).
17. Šipka, Milan (ur.). *Kompjuterska obrada lingvističkih podataka*. Sarajevo, Institut za književnost i jezik, 1978.
18. Utvić, M. *Izgradnja referentnog korpusa savremenog srpskog jezika*, doktorska teza, Filološki fakultet, Univerzitet u Beogradu, 2014.
19. Vitas, D. Mogućnosti automatske obrade teksta, *Prevodilac*, Društvo naučnih i stručnih prevodilaca RS, Beograd, sv. 4, 1982., pp. 5-15.

20. Vitas, D. (1997). O elementarnoj morfografemskoj klasi, *Naučni sastanak slavista u Vukove dane*, MSC, sveska 26/2, Beograd 1997, pp. 195-206.
21. Vitas, D. *Prevodioci i interpretatori (Uvod u teoriju i metode kompilacije programskih jezika)*. Matematički fakultet, Beograd, 2006.
22. Vitas, D. Lokalne gramatike srpskog jezika. *Zbornik Matice srpske za slavistiku 71/72, Novi Sad : Matica srpska,* 2007, стр. 305-317.
23. Vitas, D. O problemu ne(pre)poznate reči u obradi tekstova na srpskom jeziku, Zbornik Matice srpske za filologiju i lingvistiku, 50 (1-2), 2007. стр. 111-120.
24. Vitas, D.; Popović LJ.; Krstev C. i dr. *The Serbian Language in the Digital Age*, Springer-Verlag.
25. Vitas, D.; Krstev, C. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, vol. LXIII, pp. 279-292, Warszawa, 2012.
26. Vitas. D., Popović, Lj., Krstev, C., Zečević, A. How to Differentiate the Closely Related Standard Languages?, In *Proceedings of the Second International Conference Computational Linguistics in Bulgaria (CLIB 2016),* 2016, pp. 559-574.
27. Vitas, D.; Krstev, C. Restructuring Lemma in a Dictionary of Serbian. In *Zbornik 7. mednarodne multikonference "Informacijska druzba IS 2004", Jezikovne tehnologije* Ljubljana, Slovenija, Institut "Jožef Stefan", Ljubljana, 2004.

Rečnici

RSANU (1959-2021): *Речник срӣскохрвāтскоӣ књижевноӣ и народноӣ језика*, књ. 1-21, Београд

RMS (1968-1976): *Речник срӣскохрвāтскоӣа књижевноӣ језика*. Нови Сад – Загреб: Матица српска – Матица хрватска.

JRMS: Николић, М. (ур.) (2007) *Речник срӣскоӣ језика*. Нови Сад: Матица српска.

ON COMPUTER PROCESSING OF THE SERBIAN LANGUAGE

Summary

The paper presents the development of language resources for the Serbian language and the basic results achieved within the research group for language technologies, which brings together mainly researchers from the Faculty of Mathematics and the Faculty of Philology, University of Belgrade.

Initially, the key events that influenced the development of this area are presented, and then, subsequently, the basic concepts used to describe the initial steps of processing a language are presented and illustrated. The examples of some of the more complex tasks that are solved in this area have been given. The elements of the part of formal language theory – the description of regular expressions – as a means for formalizing the description of morphological processes in one language, as well as their application to the description of the morphology of Serbian, are listed in the most brief and quite informal

ways. Then, a model of electronic (morphological) dictionary intended for processing Serbian is presented, as well as its connection with other lexical databases and semantic networks (*WordNet, Prolex*). Some of the qualifiers assigned to the determinants in the e-dictionary and their purpose during language processing have been described. The basic features of the *Unitex* software system are indicated, the system, which, on the one hand, enables the generation of morphological e-dictionaries, and on the other hand, their exploitation during corpus processing. Of particular importance are the local grammars that can be formed in this system, which describe the structure of complex linguistic objects. Resources developed under this system, as well as some of its modules are part of more complex applications for processing the Serbian language.

Different corpora of the Serbian language are described, and above all, the corpus of the modern Serbian language, which has 112 million words, which is lemmatized, and information on the type of word is added to each lemma. This corpus is used primarily by researchers and students of the Serbian language. Parallel corpora in which one of the languages is Serbian are also briefly described. The French-Serbian, English-Serbian, German-Serbian, Croatian-Serbian and Serbian-Serbian corpora were developed (from multiple translations of one work into our language). The alignment was performed at the level of equivalent segments and was manually verified.

The basic outlines of some of the developed applications are shown, namely the *Bibliša* and *Leximirka* systems. *Bibliša* is a system for searching multilingual corpora in which the possibility of expanding queries via e-dictionaries and various lexical multilingual databases has been built in, while *Leximirka* is a system for developing monolingual dictionaries that relies on available machine-readable and electronic dictionaries. In conclusion, the problem of wider use of developed resources and tools in the domestic environment is indicated.