Stefan Ivanović i Kosta Jevtić

Automatizovana identifikacija prefiksa na imenicama srpskog jezika zasnovana na pravilima

Dosadašnja psiholingvistička istraživanja su ukazala na značajan robustan efekat frekvence jezičkih izraza pri kognitivnoj obradi jezika. Samim time, tokom istraživanja koja se bave morfološki složenim rečima, bitno je uzimati u obzir frekventnost pojedinačnih morfema reči. Cilj ovog rada je da se napravi program koji će biti prvi korak ka pravljenju baze frekvencija prefiksa, budući da ona ne postoji za srpski jezik. Program je pisan u programskom jeziku Python 3.6.7 i imao je za zadatak da na korpusu identifikuje sve imenice koje sa sinhronog stanovišta sadrže prefiks. U sklopu programa su definisana 4 osnovna pravila, a da bi program u imenici prepoznao prefiks, bilo je potrebno da bude zadovoljeno makar jedno pravilo. Lista imeničkih prefiksa i njihovih alomorfa kojom se program služio je napravljena na osnovu prefiksa domaćeg porekla opisanih u Tvorbi reči u savremenom srpskom jeziku Ivana Klajna. Program smo testirali na lematizovanom i anotiranom elektronskom korpusu savremenog srpskog jezika, koji sadrži 19000 imeničkih lema. Ukupno je 4300 imenica zadovoljilo neki od uslova zadatih programom. Tačnost programa je na po 100 prostim slučajnim uzorkom izabranih imenica evaluiralo troje filološki stručnih ispitanika. Na osnovu njihovih odgovora je utvrđeno da je prosečna preciznost programa (procenat tačnih identifikacija koje je program napravio) 76.3%, da prosečan odziv (šansa da program prepozna da imenica sadrži prefiks) iznosi 77.8%, a da je F1 skor $F_1 = 0.77$.

Stefan Ivanović (1999), Novi Sad, Seljačkih buna 85, 4. razred Gimnazije „Jovan Jovanović Zmaj”

Kosta Jevtić (1999), Požarevac, Prizrenska 16, 4. razred Požarevačke gimnazije

MENTORI: MA Aniko Kovač, Filozofski fakultet Univerziteta u Novom Sadu

MA Miloš Košprdić, Filozofski fakultet Univerziteta u Novom Sadu

Uvod

Prefiks je morfema koja se sastavlja sa osnovom ispred nje i tako pravi novu celinu koja ima modifikovano značenje (Stanojčić i Popović 1992: 64). Postoje različiti modeli skladištenja morfološki složenih reči u mentalnom leksikonu. Po modelu integralnih jedinica, morfološki kompleksne reči skladištene su u leksikonu kao celovite jedinice (Kostić 2006: 285), a po modelu dekompozicije u leksikonu su kao zasebne jedinice

skladištene morfeme od kojih je reč sačinjena, a ne cela reč (Taft i Forster 1975, prema Kostić 2006). Prema modelu dvostrukog puta, način skladištenja reči unutar leksikona uslovljen je njenom frekvencom (Baayen *et al*. 1997, prema Kostić 2006). U istraživanjima zasnovanim na modelima mentalnog leksikona koji podržavaju reprezentaciju pojedinačnih morfema pokazano je da frekvenca morfema od kojih se reč sastoji utiče na brzinu kojom će se ta reč obraditi (Laudanna *et al*. 1995; Alegre i Gordon 1999; Baayen *et al*. 1997, prema Kostić 2005; Colé *et al*. 1989; Wurm 1997; Ford *et al*. 2010).

Cilj ovog rada je razvoj algoritma za automatsku identifikaciju prefiksa na imenicama srpskog jezika. Algoritam za identifikaciju prefiksa kao svoj izlaz može dati podatke o distribuciji i frekvenci prefiksa u imenicama u srpskom jeziku. Budući da trenutno ne postoji baza frekventnosti prefiksa za srpski jezik, podaci dobijeni ovim algoritmom bi činili prvi korak u pravljenju takve baze. Pošto je pokazano da frekvenca jezičkih iskaza ima robustan efekat pri kognitivnoj obradi jezika, potrebno je je imati uvid u frekvence pojedinačnih morfema morfološki složenih reči tokom sprovođenja psiholingvističkih istraživanja. Podaci o zastupljenosti pojedinačnih morfema su takođe važni i za konstruisanje algoritama za segmentaciju reči na subleksičkom nivou u oblasti računarske lingvistike i obrade prirodnih jezika.

U srpskom jeziku prefiksacijom mogu nastati imenice, pridevi, glagoli i prilozi. U ovom radu ograničili smo se na imenice sa prefiksom, jer su kod njih prefiksi manje zastupljeni nego kod glagola, pa smo samim tim očekivali manji broj i veću pokrivenost tipova prefiksacije. U ovom radu ćemo se baviti imenicama koje sa sinhronog stanovišta sadrže prefiks – imenicama nastalim prefiksacijom (*nečovek* < *ne-* + *čovek*), prefiksalno-sufiksalnom tvorbom (*zatišje* < *za-* + *-tiš-* + *-je*) i deverbacijom prefigiranih glagola (*uveličavanje* < *uveličava-* (+ *-ti*) + *-nje*).

Izdvajanje prefiksa

Budući da ne postoji konačna lista prefiksa koji se javljaju na imenicama u srpskom jeziku, bilo ju je potrebno napraviti za potrebe ovog rada. Listu smo napravili na osnovu poglavlja o domaćim imeničkim prefiksima u *Slaganju i prefiksaciji* Ivana Klajna (2002: 183–194). Tokom rada na programu iz liste prefiksa smo izbacili prefiks *pa-*, zbog toga što je u savremenom srpskom jeziku u upotrebi samo jedna imenica sa ovim prefiksom (*paperje*), te je diskutabilno koliko je ovaj prefiks uopšte prepoznatljiv kao takav među govornicima srpskog jezika.

U radu na automatizovanoj identifikaciji prefiksa naišli smo na dva problema pri njihovoj segmentaciji: neki prefiksi imaju više oblika (javljaju se aloprefiksi), a neke imenice počinju slovnim nizom koji je jednak nekom od prefiksa ali nije prefiks (pseudoprefiksi).

Problemi u morfološkoj segmentaciji prefiksa u srpskom jeziku

Aloprefiksi

Budući da srpski jezik spada u flektivne jezike, između prefiksa i osnove može doći do međumorfemske fuzije u zavisnosti od fonološkog okruženja. Različiti oblici istog prefiksa nazivaju se aloprefiksi. Zbog pravopisne norme, koja je pretežno fonološka, aloprefiksi se u načinu zapisivanja mogu razlikovati od osnovnog oblika prefiksa na nekoliko načina: po zvučnosti poslednje foneme, po mestu tvorbe poslednje foneme i po broju grafema. Tako se, primera radi, prefiks *iz-* može javiti u 4 oblika: *is-* ispred bezvučnih suglasnika (*isparenje*), *iš-* ispred *č* i *ć* (*iščuđenost*), *i-* ispred *s*, *š*, *z* i *ž* (*isušenost*, *išibati*, *izračiti*, *iživeti*) i *iza-* kada se nalazi ispred reči koja počinje suglasničkom grupom (*izazvati*). Ovaj problem smo rešili proširivanjem liste prefiksa, tako da se u njoj nalaze i svi aloprefiksi.

Pseudoprefiksi

Za identifikaciju prefiksa nije bilo dovoljno da se utvrdi da li imenica počinje na slovni niz koji odgovara jednom od prefiksa, jer velik broj imenica koje počinju na slovni niz koji je formom isti nekom prefiksu u stvari nemaju prefiks. Ovakvi slovni nizovi nazivaju se pseudoprefiksi. Reči sa pseudoprefiksom su, na primer, *sako* (*sa-* kao pseudoprefiks), *privatnost* (*pri-* kao pseudoprefiks), *natrijum* (*na-* kao pseudoprefiks), *proces* (*pro-* kao pseudoprefiks). Zbog toga je potrebno da se ustanove pravila kojima bi se imenice koje počinju na prefiks razlikovale od imenica koje počinju na pseudoprefiks.

Algoritam za identifikaciju i validaciju prefiksa

Zbog navedenih specifičnosti morfološke segmentacije za srpski jezik ne postoji konačan spisak kriterijuma koje imenica treba da zadovolji da bi se moglo zaključiti da u sebi sadrži prefiks. Stoga smo do uslova koji će biti opisani došli uočavanjem pravilnosti tokom kodiranja i testiranja programa. Izdvojili smo šest hijerarhijski ustrojenih uslova. Ukoliko se tokom provere neki uslov zadovolji, program bi na toj imenici validirao prisustvo prefiksa.

1. Identifikacija kandidata prefiksa na osnovu podudarnosti početnog niza reči sa spiskom mogućih prefiksa i zadovoljenosti pravila za javljanje aloprefiksa

Ukoliko imenica počinje slovnim nizom jednakim nekom od aloprefiksa, program je proveravao da li je u datoj reči fonološko okruženje takvo, da se aloprefiks može javiti. Na primer, u reči *otcepiti* pronađen je slovni niz *ot*, što je alomorf od prefika *od-*, a pošto se ispred *c* umesto

prefiksa *od-* javlja aloprefiks *ot-*, uslov je zadovoljen. Primer gde se slovni niz koji odgovara jednom od alomorfa ne nalazi u odgovarajućem okruženju je reč *zova*. Pronađeno je slovo *z*, što je alomorf od prefiksa *s(a)-*, međutim, sledi ga samoglasnik (aloprefiks *z-* se u takvom okruženju ne može javiti); stoga, uslov za prepoznavanje aloprefiksa nije ispunjen i program prelazi na narednu reč.

2. Uslovi leksikalizovanosti kandidata za osnovu

Prvi uslov koji smo konstruisali je uslov leksikalizovanosti kandidata za osnovu. Program tim uslovom ispituje da li posle prepoznavanja i odstranjivanja kandidata za prefiks ostaje slovni niz koji je imenica srpskog jezika (npr. *(ne)čovek*). Iako takav slovni niz ne mora nužno biti tvorbena osnova reči, u ovom radu ćemo ga, budući da istorijski nastanak reči nije relevantan za predmet našeg rada, nazivati kandidatom za osnovu. Ovaj uslov je ispunilo 1587 imenica. Zbog šuma u korpusu, dodat je uslov da posle otklanjanja prefiksa dobijeni slovni niz mora biti duži od jedne grafeme.

3. Uslov produktivnosti potencijalne osnove

Budući da postoje imenice poput *pregrada* i *ograda*, koje imaju prefikse (u ovom slučaju *pre-* i *o-*) i bez tih prefiksa imaju isti ostatak reči (*-grada*) koji ne postoji kao samostalna reč, shvatili smo da uslov leksikalizovanosti nije dovoljan. Stoga smo dodali i uslov produktivnosti osnove kao drugi uslov, kojim program ispituje da li kandidat za prefiks može da se zameni nekim drugim prefiksom, a da dobijeni slovni niz postoji u listi imenica. Ukoliko postoji bar još jedna takva reč, program je tu reč prepoznao kao imenicu s prefiksom. Tako je recimo u slučaju imenice *razgrađivanje*, program prepoznao da počinje slovnim nizom koji odgovara jednom od prefiksa (*raz-*), ali ostatak reči nije imenica srpskog jezika (**građivanje*) tako da prvi uslov nije zadovoljen. Potom je program prolazio kroz listu prefiksa i pokušavao da umesto onog koji je na imenici (*raz-* u ovom slučaju) postavi neki iz liste tako da dođe do nečeg što je imenica srpskog jezika (prefiks *u-* u ovom slučaju, pa je dobijena imenica *ugrađivanje*. Ovime je imenica razgrađivanje zadovoljila drugi uslov). Kandidat za osnovu je morao biti duži od jedne grafeme da bi se ovaj uslov mogao zadovoljiti kako bi se eliminisala mogućnost da na parovima reči poput *san* i *don* budu prepoznati prefiksi (u ovom slučaju *sa-* i *do-*). Drugi uslov je ispunilo 1465 imenica.

4. Konverzija deverbala

Veliki broj deverbalnih imenica nije uspeo da zadovolji ni prvi ni drugi uslov, stoga su dodati treći i četvrti uslov. Deverbalne imenice su imenice nastale dodavanjem sufiksa na glagolsku osnovu. Ako je glagol od kojeg je deverbal izveden prefigiran, sam deverbal takođe u sebi sadrži prefiks.

Listu glagolskih sufiksa smo napravili na osnovu registra prostih i složenih imeničkih sufiksa Ivana Klajna (2003: 397–399).

Ukoliko deverbalna imenica koja počinje slovnim nizom koji bi mogao biti prefiks nije zadovoljila jedan od prva dva uslova, program je prolazio kroz listu sufiksa da bi utvrdio da li se imenica završava na slovni niz koji odgovara nekom od sufiksa iz liste. Ako se završava, taj završni niz (sufiks) se odstranjuje, tako da ostane samo kandidat za glagolsku osnovu. Zatim bi započela iteracija kroz listu nastavaka za infinitiv. Ovu listu smo sastavili na osnovu poglavlja o tvorbi glagola u *Savremenom srpskohrvatskom jeziku* Mihaila Stevanovića (1964: 603–624). Program je, dodavši nastavak za infinitiv na kandidata za glagolsku osnovu, proveravao da li je tako konstruisani slovni niz glagol srpskog jezika. Ukoliko bi program pronašao takav glagol u listi, to je bio dokaz za to da se radi o deverbalu. Ukoliko se glagol nije mogao rekonstruisati, imenica je preskakala peti korak.

5. Rekurzivna primena pravila 1 i 2 na glagole

Ako imenica počinje slovnim nizom koji odgovara nekom prefiksu i prepoznata je kao deverbal, treba videti da li je glagol od kojeg je nastala prefigiran. Način da se to utvrdi je da program proveri zadovoljava li taj glagol neki od prethodno opisana dva uslova (uslov leksikalizovanosti osnove i uslov produktivnosti kandidata za osnovu). Primer imenice koja zadovoljava treći uslov je *doživljaj* koja nije uspela da zadovolji prva dva uslova. Program je prepoznao da počinje na prefiks (*do-*) i da se završava na sufiks (*-ljaj*). Potom je pronašao nastavak za infinitiv koji može da se doda umesto sufiksa (*-eti*) tako da se dobije glagol koji se nalazi u korpusu (*doživeti*). Na kraju je ustanovio da kada se sa dobijenog glagola ukloni prefiks, rezultirajući slovni niz takođe bude leksikalizovan (*živeti*); stoga je doživljaj prepoznata kao imenica s prefiksom. Broj deverbalnih imenica čiji glagol od kojeg su izvedene zadovoljava prvi uslov je 504, a drugi 365.

U prvobitnim verzijama programa, ako posle uklanjanja prvog pronađenog slovnog niza koji odgovara nekom od sufiksa program nije mogao da rekonstruiše glagol, ustanovio je da se ne radi o imenici sa prefiksom. Međutim, moguće je da se reč završava na slovni niz koji odgovara većem broju sufiksa. Na primer, u reči *prerađevina* program bi mogao prepoznati sufikse *-evina*, *-ina* i *-a*, od kojih bi se samo skidanjem sufiksa *-evina* mogao rekonstruisati glagol. Zbog toga, program za svaku imenicu koja prolazi proveru konstruiše posebnu listu sufiksa kojima bi ona mogla da se završava. Posle bi pojedinačno iterirao kroz članove te liste i rekonstruisali bi se glagoli. Ako bi se otklanjanjem bilo kojeg sufiksa iz liste mogao rekonstruisati glagol, program bi dalje iterirao kroz već opisane provere.

6. Konverzija imenica nastalih sufiksacijom trpnog glagolskog prideva i uklanjanje posledica jotovanja

Određeni broj imenica nastalih dodavanjem derivacionog sufiksa na trpni glagolski pridev (*skamenjen* + *-ost* → *skamenjenost*) nije uspeo da zadovolji ni jedan od pređašnjih uslova. Nastavci su za potrebe programa preuzeti iz poglavlja o trpnim pridevima u *Savremenom srpskohrvatskom jeziku* Mihaila Stevanovića (1964: 347–8). Kada se sa imenice uklonio niz koji odgovara jednom od sufiksa, provereno je da li se dobijeni niz završava na neki od nastavaka za trpni glagolski pridev, potom se i taj nastavak otklanjao, a od kandidata za osnovu je program dodavanjem infinitivnih nastavaka pokušavao da rekonstruiše glagol.

Kada se nastavak za trpni glagolski pridev *-en* doda na prezentsku osnovu, može doći do jotovanja (npr. *postideti* → *postiđen*). Da bi se glagol mogao pravilno pronaći, potrebno je pravilo za otklanjanje posledica jotovanja kod ovakvih imenica. Pošto ne postoji konačna lista slovnih nizova koji na granici osnove i sufiksa kod trpnih glagolskih prideva podležu jotovanju, napravili smo je neposrednim prolaskom kroz listu ovakvih imenica koje nisu do tog momenta prošle ovaj uslov, osvrćući se na odgovarajući glagol od kojeg je ta imenica izvedena i razlog zbog kojeg nijedan uslov do tada nije zadovoljen.

7. Rekurzivna primena uslova 1 i 2 na glagole od kojih su nastale imenice nastale sufiksacijom trpnog glagolskog prideva

Analogno sa slučajem deverbala, kada bi program rekonstruisani glagol identifikovao u listi glagola, da bi se ustanovilo da li se na glagolu zaista nalazi prefiks, bilo je potrebno da zadovolji uslov leksikalizovanosti ili produktivnosti osnove.

Evaluacija algoritma

Algoritam je implementiran u programskom jeziku Python 3.6.7. Implementacija algoritma dostupna je pod licencom GNU General Public License v3.0 (Ivanović i Jevtić 2019). Kao korpus smo koristili SrpLemKor koji je podskup korpusa SrpKor i sadrži 3.7 miliona lema (Utvić 2011; Popović 2010). SrpLemKor je anotirani korpus srpskog jezika Matematičkog fakulteta Univerziteta u Beogradu sastavljen na osnovu tekstova srpskih pisaca iz XX i XXI veka, naučnih i naučno-popularnih tekstova, administrativnih tekstova i opštih tekstova (iz novina, magazina, časopisa).

Najpre smo napisali program koji iz korpusa izdvaja sve imeničke leme u zaseban fajl. U njemu se našlo 18994 imeničkih lema. Na osnovu ovog fajla, program je dalje pravio listu imenica kroz koju je program iterirao. Na isti način smo iz korpusa izdvojili sve glagolske leme i napravili listu kojom se program služio prilikom provere 3, 4, 5. i 6. uslova. U toj listi se našlo 5418 glagolskih lema.

Evaluaciju tačnosti rezultata programa izvršilo je troje ispitanika, diplomiranih filolologa srbista. Svaki ispitanik je dobio listu koja se sastoji od po 50 nasumičnih imenica na kojima je identifikovan prefiks i 50 onih na kojima nije. Njihov zadatak bio je da procene koje imenice sadrže prefiks na osnovu sinhronog stanovišta. Ispitanici nisu bili upućeni u odgovore do kojih je program došao.

Za procenu učinka programa koristili smo mere preciznosti i odziva i njihovu harmonijsku sredinu. Preciznost (engl. precision) je jedna od mera procene tačnosti, koja predstavlja procenat prepoznatih imenica s prefiksom među svim pozitivno identifikovanim imenicama. Preciznost se dobija deljenjem ukupnog broja stvarno pozitivnih imenica i zbira stvarno pozitivnih i lažno pozitivnih imenica. Odziv (engl. recall) je jedna od mera procene tačnosti, koja predstavlja procenat pozitivno identifikovanih imenica s prefiksom među svim imenicama s prefiksima. Dobija se deljenjem ukupnog broja stvarno pozitivnih imenica i zbira stvarno pozitivnih i lažno negativnih imenica.

U 76.3% slučajeva program je ispravno procenio da li se radi o imenici sa prefiksom ili ne (preciznost programa), a sa šansom od 77.8% je mogao da prepozna prefiks na imenici sa prefiksom (odziv). Harmonijska sredina preciznosti i odziva (F1 skor) iznosi $F_1 = 0.77$.

Ka frekvencijskom rečniku prefiksa – dalji koraci u unapređenju algoritma

Ovaj program je prvi korak ka pravljenju celovite baze frekvencije prefiksa u srpskom jeziku. U ovom radu smo se bavili samo imenicama, ali algoritam valja proširiti tako da zahvati i ostale vrste reči koje mogu da sadrže prefiks, te bi krajnji produkt bio celokupni frekvencijski rečnik prefiksa u srpskom jeziku.

Program je imao problema kod adekvatnog identifikovanja određenih prefiksa (*na-*, *o-*, *po-* i *pre-*) na imenicama čije osnove počinju ili čiji se prefiks završava slovom *d*, jer se u srpskom jeziku javljaju prefiksi dosta slični njima (*nad-*, *od-*, *pod-* i *pred-*). Na primer, program bi prepoznao da reč naduvavanje počine prefiksom *nad-*, a to je potvrdio pošto se može zameniti prefiksom *pred-* tako da se dobije reč srpskog jezika (*preduvavanje*), iako su u pitanju prefiksi *na-* i *pre-*. Zbog nekomplementarne distribucije ovih parova prefiksa, nismo mogli konstruisati pravilo koje bi omogućilo da program ispravno prepozna koji je prefiks u pitanju.

Nešto na šta nismo obratili pažnju, a što bi moglo dodatno povećati obuhvat imenica s prefiksima jeste uračunavanje prefiksa koji se mogu javiti na glagolima. Opisani imenički prefiksi u *Slaganju i prefiksaciji* (Klajn 2002) odnose se na prefigirane imenice, međutim, budući da se na glagolima mogu naći i neki prefiksi koji se ne nalaze na imenicama koje nisu nastale deverbacijom (poput *pro-*), bilo bi potrebno dodati i njih u listu prefiksa, upravo zbog deverbala. Takođe, prilikom provere trećeg, četvrtog, petog i šestog uslova, potrebno je ograničiti listu prefiksa prilikom rekon-

strukcije glagola tako da ne uključuju imeničke prefikse koji se na glagolima ne javljaju (poput *bez-*). Ipak, pretpostavljamo da ova odstupanja nemaju značajan uticaj na tačnost algoritma, budući da se imenički i glagolski prefiksi u najvećem broju poklapaju.

Budući da u srpskom jeziku na jednoj imenici može doći do nagomilavanja prefiksa (*onemogućavanje*, *o+ne+mogućavanje*), bilo bi poželjno dopuniti algoritam pravilom za prepoznavanje više prefiksa na jednoj reči. Podaci dobijeni takvim algoritmom mogu poslužiti za analizu distribucije prefiksa u rečima u kojima se javlja nagomilavanje prefiksa.

Tokom rada na algoritmu, pokušali smo napraviti pravilo za deverbale koji se završavaju nultim sufiksom (npr. *prevod*). Ako se dotle ni jedan uslov nije ispunio, program je, koristeći nastavke za infinitiv, konstruisao glagole za koje je proveravao postoje li u korpusu, vodeći računa o uslovu leksikalizovanosti i uslovu produktivnosti kandidata za osnovu. Tokom testiranja, broj pozitivno identifikovanih imenica bio je veoma mali, kao i njihova mera preciznosti. Budući da nismo uočili sistematske razlike između stvarno pozitivnih i lažno pozitivnih imenica koje bi ovaj uslov učinile doradivim, odlučili smo da ga zanemarimo. Detaljnijim pregledom distribucije nultog sufiksa bi u budućnosti algoritam možda mogao biti dopunjen ovim uslovom.

Podaci o frekvencama pojedinačnih prefiksa koji su dobijeni implementacijom algoritma tabelarno su prikazani u prilogu 1.

Budući da do sada nije bilo pokušaja da se napravi frekvencijski rečnik prefiksa za srpski jezik, pristup zasnovan na pravilima je bio neophodan prvi korak ka stvaranju takvog rečnika. Spisak imenica i njihovih prefiksa dobijen ovim pristupom može se iskoristiti za pothranjivanje neuronske mreže, koja bi mogla prepoznati prefikse preciznije nego pristupom zasnovanim na pravilima, i tako bi se moglo doći do verodostojnijih podataka za celokupan frekvencijski rečnik prefiksa u srpskom jeziku.

Literatura

Alegre M., Gordon P. 1999. Frequency Effects and the Representational Status of Regular Inflections. *Journal of Memory and Language*, **40** (1): 41.

Baayen R. H., Dijkstra T., Schreuder R. 1997. Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model. *Journal of Memory and Language*, **37** (1): 94.

Colé P., Beauvillain C., Segui J. 1989. On the Representation and Processing of Prefixed and Suffixed Derived Words: A Differential Frequency Effect. *Journal of Memory and Language*, **28** (1): 1.

Ford M. A., Davis M. H., Marslen-Wilson W. D. 2010. Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, **63** (1): 117.

Ivanović S., Jevtić K. 2019. Deprefiksator (Version v0.1). Zenodo. doi:10.5281/zenodo.3534243

Klajn I. 2002. *Tvorba reči u savremenom srpskom jeziku*. *Deo 1, Slaganje i prefiksacija*. Beograd: Zavod za udžbenike i nastavna sredstva

Klajn I. 2003. *Tvorba reči u savremenom srpskom jeziku*. *Deo 2, Sufiksacija i konverzija*. Beograd: Zavod za udžbenike i nastavna sredstva

Kostić A. 2006. *Kognitivna psihologija*. Beograd: Zavod za udžbenike i nastavna sredstva

Laudanna A., Burani C. 1995. Distributional properties of derivational affixes: Implications for processing. U *Morphological Aspects of Language Processing: Cross-Linguistic Perspectives* (ed. L. B. Feldman). Hillsdale: Lawrence Erlbaum Associates, str. 345–364.

Popović Z. 2010. Taggers applied on texts in Serbian. *INFOtheca*, **11** (2): 21a–38a.

Stanojčić Ž., Popović Lj. 1992. *Gramatika srpskog jezika*. Beograd: Zavod za udžbenike i nastavna sredstva

Stevanović M. 1964. *Savremeni srpskohrvatski jezik (gramatički sistemi i književnojezička norma): Uvod, fonetika, morfologija*. Beograd: Naučno delo

Utvić M. 2011. Annotating the Corpus of contemporary Serbian. *INFOtheca*, **12** (2): 36a.

Wurm L. H. 1997. Auditory processing of prefixed English words is both continuous and decompositional. *Journal of memory and language*, **37** (3): 438.

Stefan Ivanović and Kosta Jevtić

A Rule-Based Approach to the Automatic Identification of Prefixes in Serbian Nouns

Psycholinguistic research has shown that the frequency of linguistic units has a robust effect on cognitive language processing. For that reason, when conducting research which deals with morphologically complex words, it is important to take into account the frequencies of each of their individual morphemes. Our goal was to create a program which would be the first step towards creating a prefix frequency database, since one does not yet exist for the Serbian language. The program was written in the Python programming language, version 3.6.7, and its task was to identify all the nouns in the corpus that contain a prefix from a synchronic point of

view. There were 4 rules defined within the code, and in order for the program to recognize a prefix within a noun, at least one of them had to be met. The program used a list of noun prefixes and their allomorphs based on the prefixes of Slavic origin described in *Tvorba reči u savremenom srpskom jeziku* by Ivan Klajn. The program was tested on a lemmatized and annotated electronic corpus of contemporary Serbian language, which contains 19000 noun lemmas. In total, 4300 nouns met at least one of the requirements defined in the program. The accuracy of the program was evaluated by 3 philologists each of whom was given a simple random sample of 100 nouns. Based on their responses, it was concluded that the average precision of the program (the percentage of correct identifications made by the program) was 76.3%, the average recall (the possibility that a prefix on a noun would be identified as such) was 77.8% and the F1 score was 0.77.

Prilog: Frekvencijski rečnik imeničkih prefiksa prema podacima dobijenim implementacijom algoritma

Prefiks	Frekvenca	Prefiks	Frekvenca
po	411	*sa*	116
s	397	*ob*	95
u	328	*uz*	66
o	279	*pred*	64
ne	262	*su*	63
za	259	*samo*	51
iz	251	*nad*	38
pre	238	*pra*	37
na	192	*polu*	34
od	172	*bez*	20
pri	164	*protiv*	13
raz	156	*među*	9
do	135	*mimo*	1
pod	122	*van*	1