Koliko puta kažeš Vučić

Po analizi urađenoj za potrebe ovog teksta, najfrekventnija imenica u "Vremenu" tokom poslednjih deset godina je "godina", Srbija je prva među državama, a ove godine najčešće pominjana vlastita imena su Aleksandar Vučić i Ana Brnabić

Ove godine navršava se 15 godina od kako je na veb postavljen korpus savremenog srpskog jezika (http://www.korpus.matf.bg.ac.rs/), a 25 od početka rada na razvoju elektronskog rečnika srpskog. Ova dva važna jezička resursa predstavljaju deo osnovnog i nezaobilaznog inventara potrebnog jednom jeziku kako bi se održao u informatičko doba. Njihova primena je dvostruka. S jedne strane, oni obezbeđuju empirijsku građu neophodnu u lingvističkim i leksikografskim istraživanjima, a sa druge, pružaju potrebne podatke za razvoj računarskih programa koji obrađuju tekstove na srpskom jeziku.

Prva verzija korpusa iz 2003. godine sadržavala je oko 23 miliona reči, dok ih u današnjoj verziji ima oko 122 miliona. Ove reči su uključene sa svojim kontekstom i osnovnim bibliografskim podacima, a grupisane su po funkcionalnim stilovima. Svakoj reči je dodeljena i lema (odnosno uobičajeni rečnički oblik) i pridružen podatak o vrsti reči. Korpus koristi više od 700 domaćih i inostranih istraživača i studenta.

ŠTA JE E-REČNIK

Činjenica da je korpus u elektronskom obliku, omogućava da se na osnovu njegovog sadržaja automatski sračunaju frekvencije upotrebljenih reči i njihovih lema. Sračunati su i statistički jezički modeli, a to su frekvencije uzastopnih pojavljivanja dve, tri i više reči. Ovi podaci se koriste, pre svega, za usmeravanje procesa obrade u statističkim prilazima obradi jezika. Računanje frekvencija je među najjednostavnijim zadacima koji se postavljaju u automatskoj obradi teksta. Za složenije obrade i pouzdaniji opis samog jezika neophodno je raspolagati formalizovanim opisom i preciznom klasifikacijom svojstava reči. Takav opis pruža elektronski rečnik (skr. e-rečnik) koji, pored drugih podataka, može sadržavati i frekvenciju reči.

E-rečnik je resurs koji je namenjen programima za obradu teksta na prirodnom jeziku. Naime, u digitalnom tekstu u kome čitalac prepoznaje lako reči i rekonstruiše njihovo značenje, računar "ne vidi" jezički organizovan materijal. Kao sredstvo da se rečima dodele jezički relevantne informacije koristi se ova vrsta rečnika.

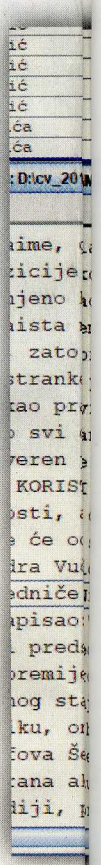
Njihov najjednostavniji oblik su oni rečnici koje koriste programi za otkrivanje tipografskih grešaka (engl. spell checker). Oni se sastoje obično iz obimnog spiska reči, bez drugih informacija, a provera se sastoji u sravnjivanju reči iz teksta sa rečima iz rečnika. Kada u rečniku nema neke reči iz teksta, signalizira se greška i moguća ispravka. Nešto složeniji primer je onaj e-rečnik koji je ugrađen u elektronske knjige kao što je Kindl ("Vreme", 1091). Tu čitalac može pozvati u pomoć tradicionalni rečnik za reč koju ne poznaje. Ova operacija je složenija od prethodne: Kindl treba da poveže oblik reči iz teksta sa odgovarajućom odrednicom u rečniku. Sličan problem, ali u suprotnom smeru, predstavlja pretraga gde za zadatu reč (obično u obliku rečničke odrednice) treba pronaći sva pojavljivanja njenih različitih oblika. Na primer, pretraga arhive "Vremena" na vebu sa odrednicom "kompaktan" morala bi pronaći i oblik "najkompaktniji" iako se on pojavio samo jednom u poslednjoj deceniji ("Vreme", 1162).

Za još složenije obrade teksta potrebni su e-rečnici koji ne uspostavljaju samo vezu između leme i njenih oblika, već opisuju odnos leme i oblika u gramatičkim kategorijama, kao i leksička svojstva same leme. Na primer, lema *crn* treba tada da bude opisana kao pridev koji označava boju, a oblik *najcrnji* kao jedan od oblika superlativa ovog prideva. Ako bismo sada poželeli da pronađemo jednim upitom u arhivi "Vremena" sve superlative prideva koji tokom poslednje decenije opisuju boje, onda bismo morali imati e-rečnik u kome su leme prideva koji označavaju boje povezane sa oblicima njihovog superlativa. Sa takvim rečnikom se tada pronalazi da su od 2010. u "Vremenu" upotrebljena 53 superlativa ovakvih prideva, da je najčešći oblik *najcrnje*, ali da se javljaju još i oblici superlativa prideva *zelen*, *žut* i *crven*.

OPISMENJAVANJE RAČUNARA

Dalje komplikacije u opismenjavanju računara dolaze od izraza čija se gramatička funkcija ili značenje ne mogu izračunati iz reči od kojih se on sastoji. Na primer, *crna ovca* nije ni crna, ni ovca, već osoba sa određenim osobinama, a promena ovog izraza po padežima nije prosti zbir mogućih oblika prideva *crn* i imenice *ovca*. Slično, izraz *s vremena na vreme* ima u tekstu prilošku funkciju koja se ne može sračunati iz njegovih delova. Dakle, za iscrpnu obradu teksta potrebno je sastaviti i popis ovakvih izraza i opisati njihove gramatičke i leksičke specifičnosti.

Ovde nije kraj onoga što će računarima praviti problem u analizi teksta. Čitave klase složenih izraza pripadaju pre enciklopedijskom nego rečničkom inventaru reči. Ovo se posebno odnosi na vlastita imena, datume, različite mere, skraćenice i druge slične reči. Potrebno je i njih popisati i dodeliti im odgovarajuće atribute. Na primer, reč *Beograd* treba u reč-



niku da ima naznačeno, pored gramatičkih svojstava, i da je u pitanju vlastito ime za grad koji se nalazi u Srbiji. Sa rečnikom u kome su na ovaj način obeleženi toponimi, mogu se automatski izdvojiti iz arhive svi tekstova u kojima se pominje bilo koje mesto u Srbiji. A ako poželimo da identifikujemo sve datume u tekstu, potrebno je opisati čitav jedan mali podjezik u kome će se, pored reči, naći cifre i razni interpunkcijski znaci. Značaj prepoznavanja ovakvih reči dolazi do izražaja kada je

Ilustracija: VREME/IH

potrebno obeležiti za dokumentarističke potrebe tekstove o tome ko je kada i gde nešto rekao ili učinio.

Sve ove operacije i mnoge druge mogu se izvršiti nad tekstom na srpskom jeziku pomoću e-rečnika čiji razvoj vodi prof. Cvetana Krstev sa Katedre za bibliotekarstvo i informatiku na Filološkom fakultetu. Sam rečnik sadrži oko 160.000 lema i više od 6.000.000 njihovih oblika, a razvija se i dalje. Model i metodologiju njegove konstrukcije razvijao je počev od 1988. francuski lingvista Moris Gros u okviru teorije leksikon-gramatike. Pod njegovim uticajem, u istom formatu izgrađeni su rečnici za više evropskih jezika, a od slovenskih, pored srpskog postoje i rečnici ruskog, bugarskog i poljskog. Sa ovakvim rečnikom moguće je rešiti sve gore navedene probleme u obradi teksta počev od pronalaženja tipografskih grešaka, otkrivanje svih oblika jedne leme prostim upitom, ali i otkrivanje složenih objekata u tekstu kao što su polileksemske jedinice, izrazi datuma ili nizovi vlastitih imena.

E-rečnici se koriste kao deo različitih informatičkih aplikacija ili kroz sistem Unitex (http://unitexgramlab.org/), slobodni softver inicijalno razvijen na Univerzitetu Pariz-Istok, a u čijem su razvoju učestvovali i beogradski matematičari-informatičari. Ovaj sistem je platforma za obradu korpusa koja se zasniva na eksploataciji sadržaja e-rečnika i koristi se prvenstveno u različitim filološkim istraživanjima i u izgradnji formalizovanih modela jezika.

IZ ARHIVE "VREMENA"

Kao primer primene e-rečnika srpskog i Unitex-a, za potrebe ovog teksta formiran je korpus hronološki uređenih članaka iz arhive "Vremena" od 2010. naovamo. Ovaj korpus se sastoji od oko 13 miliona reči, a sastavljen je od oko 422.000 različitih oblika reči. Od ovih različitih oblika reči, oni koji su se pojavili samo jednom čine 44 odsto svih upotrebljenih oblika. I pored obimnosti e-rečnika srpskog, u njemu se ne nalazi oko 46.000 oblika, a to su ili reči iz drugih jezika (warehouse, dromedarius...), tipografske greške ili nove i neobične reči (kalakurdija, kakotrebajuće, webdžije...) kojih nema ni u rečnicima Matice srpske. Ipak, u odnosu na ukupan broj pojavljivanja reči u ovom korpusu, ove "nepoznate" reči čine manje od 1 odsto.

Najčešće reči u "Vremenu" su iste kao i u korpusu savremenog srpskog jezika sa 122 miliona reči. To su *i, je, da, u, se*... Najčešćih 10 reči čini 18,43 odsto svih upotrebljenih reči u korpusu "Vremena", a najčešćih 100 – 32,89 odsto. To znači da »

je petina svih napisa od 2010. do avgusta 2018. sastavljena sa svega 10 reči, a trećina – sa svega 100 reči. Ali ovo nije svojstvo samo tekstova u "Vremenu", već svojstvo raspodele učestanosti reči u srpskom jeziku: skoro istih 10 reči su najčešće i u Andrićevoj *Travničkoj hronici* gde čine oko 20 odsto teksta. U sloju visokofrekventnih reči u "Vremenu" nalazi se i poneka imenica kao što je *godina, Srbija, ljudi, dan, vreme* ili glagoli *moći, kazati, trebati.*

Same frekvencije izolovanih reči ne pokazuju kako su one upotrebljene, niti kako se propagiraju kroz korpus. Na primer, najčešća imenica *godina* javlja se 57.240 puta što čini 0,44 odsto svih reči u korpusu. Koristeći se svojstvima e-rečnika, mogu se potražiti izrazi koji sadrže reč *godina*, a koji opisuju određeni datum ili period. Takvi izrazi će se u korpusu pojaviti 15.506 puta, a popunjavaju korpus u istom procentu kao i sama imenica *godina* jer apsorbuju veći broj reči. Primeri takvih izraza su u sekvencama "trista pedeset i četiri godine", "od 7. septembra 2017. do 14. januara 2018. godine" ili "1,2 miliona godina".

Reči koje nisu u tradicionalnim rečnicima, kao što su vlastita imena ili skraćenice različitih organizacija, čine, po pravilu, oko 12 odsto novinskih korpusa nezavisno od jezika, a ovo potvrđuje i korpus "Vremena". Najčešće vlastito ime u ovom korpusu je Srbija sa 42.268 pojavljivanja. Od imena država po učestalosti slede SAD, Nemačka, Hrvatska, Jugoslavija... a od toponima unutar Srbije najčešće se javlja Beograd (14.677), Kosovo (6.305), Vojvodina (1.830), Novi Sad (1.684), Priština (1.071), Niš (850), Kragujevac (676)... Razni toponimi Srbije pominju se ukupno oko 100.000 puta.

Reči gube ili dobijaju na značaju u funkciji vremena. Reč *Tviter* se u "Vremenu" prvi put javlja 2009. godine i u korpusu koji se sastoji od tekstova iz ove godine javlja se svega dva puta. Tokom ove decenije, tviter dobija na značaju, ali ne samo kao određeni servis na vebu. Uz ovu reč se razvila i cela porodica reči: tvit, tviteraš, tviterašica, tviterašenje, tvitovanje, tviterašenje, zatim glagoli tvitovati, tvitnuti, tvitati i retvitovati, pridevi tviteraški i tviterski itd. Na ovom primeru vidimo kako jedna reč koja se pojavila 2009. ne samo da je dobila na značaju već je razvila i niz drugih povezanih reči. Sličan fenomen je zabeležila i reč *blog*, mada ne tako plodan (bloger, blogerka, blogovanje, blogovati, mikroblog, blogerski, blogovski). S druge strane, podjednako popularni servisi na mrežama, kao što su skajp ili viber, ostali su ograničeni na svoj osnovni oblik premda nema jezičkih prepreka za konstrukciju sličnih familija reči. Za razliku od tvitera koji se širi, nacionalni *ćevapčić*, koji se u tom obliku razlio po celom svetu, u savremenom srpskom jeziku nestaje, a to potvrđuje i korpus "Vremena". Ova se reč javlja 4 puta 2010, 2 puta 2013, još jednom 2014, a onda nestaje. Zamenio ju je *ćevap* koji je, pak, promenio svoje nekadašnje značenje.

Ovi primeri nam pokazuju da se sudbina reči ne može predvideti. A tu nailazimo i na leksikografski izazov opisivanja novonastale leksike. Za neke jezike, e-rečnik se koristi svakodnevno da bi se "ulovile" nove reči, pre svega, u novinskim tekstovima. Ovo otvara zanimljive mogućnosti za različite leksikografske projekte kao što su rečnici novih ili "divljih" reči.

Kako su u e-rečniku opisana i vlastita imena, moguće je pratiti njihovu popularnost tokom vremena. Ako se potraže u korpusima "Vremena" po godištima najčešće pominjane muške i ženske ličnosti imenom i prezimenom, onda se za poslednjih 18 godina slika "popularnosti" veoma menja. Godine 2001. najčešće muško ime je Slobodan Milošević, a žensko Karla del Ponte. Pet godina kasnije najčešća su Zoran Đinđić i Angela Merkel, dok su 2010. to Boris Tadić i Hilari Klinton. Imena Aleksandra Vučića i Angele Merkel nalaze se na vrhu liste 2015, dok je u prvih 8 meseci ove godine, prvo mesto zadržao Vučić, dok se kao žensko ime prvi put na čelu javlja Ana Brnabić. Ne treba ni napominjati da se najčešće žensko ime i prezime javlja 4-5 puta ređe od najčešćeg muškog, kao i da je broj pomenutih ženskih osoba znatno manji od muških. Zanimljivo je da ime prethodnog predsednika Srbije Nikolića nikada nije došlo na vrh liste. U vreme izbora 2012. i dalje je Tadićevo ime na vrhu prema broju pojavljivanja, a već 2013. to mesto preuzima do tada malo zastupljeno ime Vučića.

Opisani sistem e-rečnika srpskog koristi se u različitim informatičkim aplikacijama. Neke od njih su teme odbranjenih doktorskih disertacija: analiza uzajamnih referenci u zakonima republike Srbije (Nebojša Vasiljević), anonimizacija korpusa koja se sastoji u zameni stvarnih izmišljenim imenima što je od značaja u analizi medicinske dokumentacije (Jelena Jaćimović), u transformaciji starinskih kuvarskih tekstova na savremeni format recepata (Staša Vujičić-Stanković), u određivanju polariteta stavova i raspoloženja u velikim kolekcijama tekstova (Miljana Mladenović)... ili objavljenih radova o analizi zastupljenosti ženskih imena u medijima (Cvetana Krstev, Sandra Gucul, Miloš Utvić i Jelena Jaćimović) ili o razlikama i sličnostima između srpskog i hrvatskog jezičkog uzusa (Ljubomir Popović, Cvetana Krstev, Anđelka Zečevič i Dušan Vitas). Takođe, ovi rečnici su osnova i za veb-aplikacije kao što je pretraga dvojezičnih tekstova (Ranka Stanković) ili veb-servisa koji su u pripremi za restauraciju dijakritičkih znakova u tekstu sa "ošišanom" latinicom i za korekturu o-ce-er-ovanih tekstova ili pretragu korpusa starijih pisaca. ¶

DUŠKO VITAS,
profesor Matematičkog fakulteta