RASPRAVE.
Časopis Instituta za hrvatski jezik i jezikoslovlje 41/2 (2015.)

UDK 811.163.42'322.2
81'322.2
Izvorni znanstveni rad
Rukopis primljen 7. IV. 2015.
Prihvaćen za tisak 9. X. 2015.

Ivan Pandžić
Institut za hrvatski jezik i jezikoslovlje
Ulica Republike Austrije 16, HR-10000 Zagreb
ipandzic@ihjj.hr

OBLIKOVANJE KORJENOVATELJA ZA HRVATSKI JEZIK

U radu je prikazana izrada dvaju korjenovateljā za hrvatski jezik (k2 i k3) koji upotrebljavaju tvorbene nastavke imenica, pridjeva i glagola kako bi odredili osnove pojavnica. Pretpostavku da će navedeni korjenovatelji postići bolje rezultate od drugih sličnih korjenovatelja za hrvatski jezik provjerili smo usporedbom njihovih preciznosti, odziva i F1-mjera s istim vrijednostima početnoga korjenovatelja (k1). U tu svrhu upotrijebljen je ručno provjereni korpus od 9775 pojavnica s određenim lemama i morfosintaktičkim oznakama. U radu su također obrađeni problemi povezani s nazivljem koje se upotrebljava u području korjenovanja.

1. Uvod

Većina se podataka čuva u tekstu. Kako bismo tim podatcima brzo i jednostavno pristupali, moramo razvijati tehnologije temeljene na znanjima iz računalstva i jezikoslovlja. Upravo je jedna od takvih tehnologija morfološka normalizacija[1]. Dva su osnovna pristupa morfološkoj normalizaciji: lematizacija i korjenovanje. Lematizacija je postupak kojim se pronalazi kanonski oblik riječi, tj. natuknički oblik u rječniku ili *lema*. Lematizator bi oblike članstvo, *članstva*, *članstava*, *članstvima* trebao svesti na lemu *članstvo*. Korjenovanje je pak postupak kojim se uklanjaju afiksi iz različitih oblika riječi kako bi našao „korijen" zajednički svim oblicima. Tako dobiven „korijen" češće je zapra-

[1] Potvrdu o učinkovitosti ovakvih postupaka možemo pronaći npr. u Šarić i dr. (2005).

vo jednak osnovi riječi, pa se neki autori koriste nazivom *pseudokorijen* (Šnajder 2010: 8). I mi ćemo se koristiti tim nazivom u ovome radu. Treba istaknuti da korjenovanje može obuhvatiti i flektivnu i derivacijsku morfološku varijaciju, tako da može dovesti do različitih stupnjeva normalizacije. Ako korjenovanje obuhvaća samo flektivnu morfološku varijaciju, govorimo o slabome (konzervativnome) korjenovanju, a ako obuhvaća i derivacijsku morfološku varijaciju, govorimo o jakome (agresivnome) korjenovanju. Postupak bi slaboga korjenovanja oblike *članstvo*, *članstva*, *članstava* i *članstvima* trebao svesti na pseudokorijen *članstv*, a postupak jakoga korjenovanja iste bi oblike trebao svesti na pseudokorijen *član*.

U ovome ćemo radu predstaviti izradu dvaju slabih ili konzervativnih korjenovatelja temeljenih na pravilima (k2 i k3), koji s pomoću tvorbenih nastavaka za imenice, pridjeve i glagole određuju zajedničke pseudokorijene pojavnica. Naša je pretpostavka da takav tip korjenovatelja može dati bolje rezultate od korjenovatelja temeljenih na pravilima koji ne sadržavaju podatke o tvorbenim nastavcima. Koristeći se obrađenim uzorkom za provjeru, izračunat ćemo preciznost, odziv i F1-mjeru korjenovatelja k2 i k3 i usporediti ih s istim vrijednostima korjenovatelja k1. U drugome je poglavlju rada predstavljena podjela morfološke normalizacije. U trećemu je poglavlju dan pregled već izrađenih korjenovatelja za hrvatski jezik, a u četvrtome izrada korjenovatelja k2, prve inačice poboljšanoga korjenovatelja koja se koristi razrađenim pravilima korjenovanja za imenice i pridjeve i pravilima korjenovanja za jedan dio glagola. U petome je poglavlju opisano vrednovanje korjenovatelja k2, a u šestome izrada korjenovatelja k3, tj. optimizirane inačice korjenovatelja k2 koji se koristi i proširenim pravilima za glagole. U sedmome je poglavlju dan pregled nazivlja korjenovanja. U osmome je poglavlju predstavljena analiza pogreška, a u devetome zaključak.

2. Podjela morfološke normalizacije

Pristupe morfološkoj normalizaciji možemo podijeliti na rječničku lematizaciju, korjenovanje temeljeno na pravilima, hibridno korjenovanje, lematizaciju temeljenu na metodama nadziranoga učenja i nenadzirane pristupe korjenovanju. Rječnička lematizacija podrazumijeva korištenje gotovih morfoloških leksikona koji oblike neke riječi povezuju s odgovarajućom lemom. Korjenovanje temeljeno na pravilima čine algoritmi temeljeni na pravilima koji pronalaze pseudokorijen riječi primjenom niza ručno kodiranih pravila odsijeca-

nja afikasa (najčešće sufikasa[2]). Hibridno korjenovanje također upotrebljava pravila odsijecanja sufikasa, ali smanjuje broj pogrešaka korištenjem rječnika. Za morfološki složene jezike prikladne su metode nadziranoga učenja za automatsku indukciju lematizacijskih pravila iz postojećih morfoloških leksikona ili morfološki označenih korpusa. Nedostatci su ovoga pristupa zahtjevna izrada velikoga broja lematizacijskih pravila i potreba za morfološkim leksikonom, vrlo skupim jezičnim resursom. Te nedostatke zaobilaze nenadzirani pristupi korjenovanju kojima se rječnici ili pravila korjenovanja induciraju automatski iz neoznačenih korpusa (Šnajder 2010: 14).

Najraširenije od navedenih pristupa jest korjenovanje temeljeno na pravilima. Prednost ove vrste morfološke normalizacije u prvome je redu njegova jednostavnost i brzina izrade. Za razliku od ostalih pristupa morfološke normalizacije, za korjenovanje temeljeno na pravilima nisu potrebni ni leksički resursi kao što su morfološki leksikoni ni specijalizirano znanje o strojnome učenju. Porterov je algoritam jedan od najučinkovitijih i najčešće korištenih algoritama za korjenovanje engleskoga jezika (Manning 2008: 31). Navedeni se algoritam sastoji od pet stadija koji se odvijaju jedan za drugim. U prvome se stadiju odvija flektivna, a u ostala četiri stadija derivacijska normalizacija. Ovo su primjeri pravila iz prvoga stadija:

SSES → SS
IES → I
SS → SS
S →

Prvo od navedenih pravila određuje da pojavnice koje završavaju nizom *sses* moraju biti skraćene za posljednja dva grafema (npr. *caresses* → *caress*), a treće određuje da pojavnice koje završavaju nizom *ss* ostaju nepromijenjene (npr. *caress* → *caress*). Ta pravila uvjetuju samo završetak leksema, ali neka druga pravila Porterova algoritma uvjetuju broj slogova osnove riječi ili prisutnost određenoga grafema. Postoje sličnosti između Porterova algoritma i korjenovatelja napravljenih za hrvatski jezik, ali morfološka složenost hrvatskoga jezika ipak zahtijeva nešto drukčiji pristup.

[2] Nije slučajno da su najčešći afiksi upravo sufiksi. Istraživanja su pokazala da su početci riječi perceptivno i kognitivno važniji od njihovih završetaka (Marković 2012: 56). Iako postoje jezici u kojima se prefiksi češće upotrebljavaju za derivaciju i fleksiju (npr. tajski), sufiksi su mnogo češći u većini jezika svijeta, kao što su i u hrvatskome jeziku. Tako su, na primjer, u riječima *trčati*, *trčim*, *trčiš*, *trkač* i *trkački* svi afiksi sufiksalni. Također, opće je pravilo u hrvatskome jeziku da derivacijski sufiksi prethode flektivnomu u istoj riječi, iako postoje iznimke (npr. *k-**oga**-god* u odnosu na *k-**omu**-god*).

3. Korjenovatelji za hrvatski jezik

Ljubešić i dr. (2007) predstavljaju dostupna rješenja za morfološku normalizaciju hrvatskoga jezika. Do navedene godine izrađena su samo dva korjenovatelja za hrvatski jezik korištena za pretraživanje informacija. Prvi je korjenovatelj za hrvatski jezik osmislila dr. sc. Tomislava Lauc napisavši pravila za imeničke i pridjevske paradigme (v. Lauc i dr. 1998). Iako je taj korjenovatelj postigao preciznost od 90,26 %, on nije bio testiran koristeći se korpusom, nego na leksikonu, ne uzimajući u obzir frekvenciju pojedinih oblika leksema. Budući da se neki padeži pojavljuju češće od drugih, to je moglo utjecati na rezultate.[3] Drugi korjenovatelj, čija je glavna svrha bila pretraživanje tekstova Narodnih novina Republike Hrvatske, napravio je Dobrica Pavlinušić (Ljubešić i dr. 2007: 314). Nažalost, taj korjenovatelj nije bio kvantitativno ispitan, pa ne možemo znati koliko je bio učinkovit.[4] U istome su članku Ljubešić i suradnici predstavili svoj, javno dostupan, agresivan korjenovatelj za hrvatski jezik temeljen na pravilima. Upravo ćemo ovaj korjenovatelj pokušati unaprijediti kasnije u radu. Za oblikovanje toga korjenovatelja korišten je novinski korpus *Vjesnik on-line*, a popis pravila za korjenovanje ručno je oblikovan na temelju 1000 najčestotnijih imenica u hrvatskome jeziku.

Šnajder i Dalbelo Bašić (2009) opisuju korjenovanje temeljeno na udaljenosti znakovnih nizova, tj. izradu korjenovatelja za hrvatski jezik koji korijenske klase izgrađuje na temelju sličnosti parova ili n-grama grafema pojavnica. Zanimljivo je da se ova vrsta korjenovatelja ne koristi podatcima o hrvatskome jeziku, tako da se može primijeniti i na druge jezike.

4. Izrada korjenovatelja k2

4.1. Temelji

U ovome ćemo poglavlju predstaviti slab ili konzervativan korjenovatelj temeljen na pravilima koji s pomoću tvorbenih nastavaka za imenice, pridjeve i glagole određuje zajedničke pseudokorijene pojavnica. Naš se korjenovatelj sastoji od nekoliko dijelova – glavnoga programa[5], pravila korjenovanja, napisanih s pomoću regularnih izraza, i transformacija. Navedena su pravila napravljena na temelju pravila korjenovanja[6] koja je oblikovao dr. sc. Nikola Ljubešić:

[3] Prema novim istraživanjima, nominativ je najčešći padež, a slijede ga akuzativ i genitiv. Ta tri padeža zajedno čine oko osamdeset posto teksta. Vokativ je najrjeđi, a ostala tri padeža mijenjaju redoslijed ovisno o vrsti teksta (Kolaković 2007: 1).

[4] Nijedan od navedenih dvaju korjenovatelja nije dostupan javnosti.

[5] Glavni je program napisao dr. sc. Nikola Ljubešić koristeći se programskim jezikom Python.

[6] Taj ćemo korjenovatelj u ovome radu zvati korjenovatelj k1.

.+[^aeiou] skoga|skima|skom|skoj|skog|skim|skih|noga|sku|sko|ski|ske|ska|nom|noj|nog|nim|nih|na|nu|no|ni|ne

.+ anjima|enjima|stvima|ovima|evima|enoga|anoga|anjem|enjem|stvom|stvo|stva|stvu

.+ anje|enje|anja|enja|enom|enoj|enog|enim|enih|anom|anoj|anog|anim|anih|eno|ano|ovi|ova|oga|ima|evi|eva |ove|eve|enu|eni|ene|anu|ani|ane|ena|ana|ama

.+ om|og|im|ih|em|oj|u|o|i|e|a[7]

Navedena se pravila primjenjuju po redu dok se jedno od tih pravila ne primijeni uspješno. Korjenovatelj koji se koristi tim pravilima, korjenovatelj k1, klasificiran je kao jak ili agresivan korjenovatelj zato što su pravila oblikovana tako da na istu osnovu mogu svesti oblike različitih vrsta riječi koje su semantički ili tvorbeno povezane, npr. *bankarstvo*, *bankara* i *bankarstva* na *bankar* ili *bogatima*, *bogatoj*, *bogatog*, *bogatstvo* i *bogatih* na *bogat*. Iako je korjenovatelj k1 vrlo učinkovit za tako malen broj pravila, njegov ustroj ipak dovodi do određenih pogrešaka. U ovome je isječku rezultata korjenovanja korpusa Index_1000, korpusa koji sadržava tisuću članaka novinskoga portala index.hr, korjenovatelj k1 pojavnici *županje* pogrešno odredio pseudokorijen *žup*, pojavnicama *župana* i *županom* pogrešno je određen isti taj pseudokorijen. Oblicima *Županjac* i *Županjaca* pogrešno je određen pseudokorijen *županjac* (umjesto *županjc*). Upravo ćemo ovakve pogreške pokušati ispraviti povećanjem broja pravila korjenovanja.

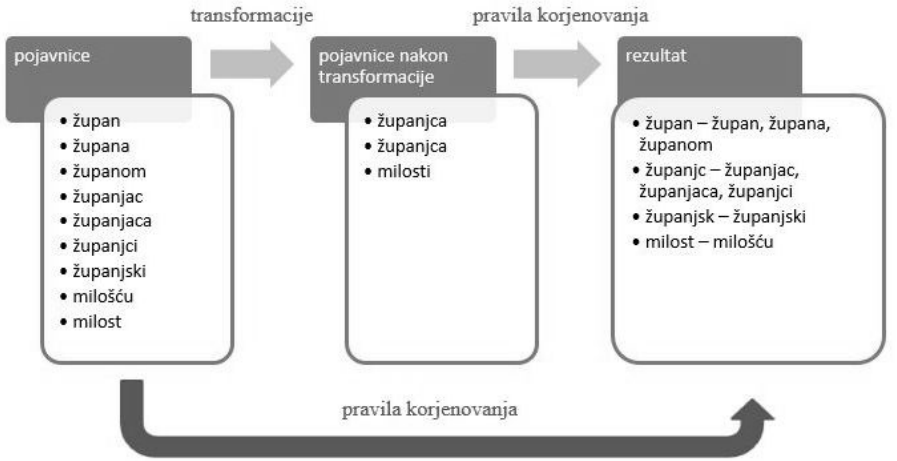
pseudokorijen	pojavnice svedene na isti pseudokorijen (pogreške u kurzivu)
žup	župni, župnog, župnim, župa, župe, župi, *župana*, *županom*, *županje*
župan	župan
županij	županijsko, županijskim, županijskoga, županiji, županijski, županijskih…
županj	županjski, županju, županji
županjac	*županjac*, *županjaca*
županjc	županjci, županjce

Tablica 1. Primjeri korjenovanja korjenovateljem k1

[7] Pravilo „.+ om|og|im|ih|em|oj|u|o|i|e|a” mogli bismo pročitati na sljedeći način: „Pronađi sve nizove znakova koji sadržavaju barem jedan znak iza kojega slijede nizovi znakova ‘om’, ‘og’, ‘im’, ‘ih’, ‘em’, ‘oj’, ‘u’, ‘o’, ‘i’, ‘e’ ili ‘a’”.

4.2. Opis postupka korjenovanja

Nakon učitavanja korpusa, program tokenizira tekst, tj. razdvaja pojavnice i sva velika slova pretvara u mala. Transformacije se primjenjuju na pojavnice s odgovarajućim završetcima. Na te se pojavnice zatim primjenjuju pravila korjenovanja s dodatnim uvjetom da svaka osnova mora sadržavati barem jedan samoglasnik ili slogotvorno r. Pravila se korjenovanja odmah primjenjuju na pojavnice kojima nisu potrebne transformacije.



Slika 1. Prikaz postupka korjenovanja

4.3. Izrada pravila konzervativnoga korjenovatelja

Već spomenutim četirima pravilima dodali smo 297 dodatnih pravila, tj. 185 za korjenovanje glagola, 62 za korjenovanje pridjeva i 50 za korjenovanje glagola. Naša je pretpostavka bila da će dodatna pravila poboljšati postupak korjenovanja. Prvo su izrađena pravila za imenice prema popisu imeničkih sufikasa iz *Tvorbe riječi u hrvatskome književnome jeziku* autora Stjepana Babića (2002: 70–72).[8] Nakon pravila za imenice izrađena su pravila za pridjeve[9] i gla-

[8] Detaljna klasifikacija uzoraka promjena, tj. mogućih tvorbenih završetaka flektivnih osnova za hrvatski jezik, dostupna je i u Tadić (1994).

[9] Budući da su fokus našega korjenovatelja bile imenice i pridjevi, dvije vrste riječi koje su najznačajnije za pretraživanje informacija, pravila za glagole rađena su samo za one glagole čiji su oblici mogli pokvariti rezultate korjenovanja imenica i pridjeva.

gole s pomoću popisa pridjevskih sufikasa iz već navedenoga gramatičkog priručnika (Babić 2002: 381–382). Ovo su primjeri dodatnih pravila za imenice, pridjeve i glagole:

.+an ima|om|a|e|i|o|u|

.+sk ijima|ijega|ijemu|ijem|ijim|ijih|ijoj|ijeg|iji|ije|ija|oga|ome|omu|ima|og|om|im|ih|oj|a|e|i|o|u

.+inj asmo|aste|ati|emo|ete|ali|ala|alo|ali|ale|aše|ahu|em|eš|at|ah|ao

Prvo navedeno pravilo svodi različite oblike imenice koja je tvorena tvorbenim sufiksom *-an*[10] na zajednički pseudokorijen, tj. na zajedničku osnovu. Prema tome pravilu, pojavnica može sadržavati bilo koje grafeme prije morfema *-an*, ali mora sadržavati taj tvorbeni morfem i jedan od sljedećih oblićnih nastavaka: *-ima*, *-om*, *-a*, *-e*, *-i*, *-o*, *-u* ili *-ø*[11]. S pomoću toga pravila pseudokorijen genitiva jednine leksema *župan* (*župana*) bit će određen kao *župan* umjesto *žup*. Drugo navedeno pravilo trebalo bi svesti različite oblike pridjeva tvorenih tvorbenim sufiksom *-ski* na zajednički pseudokorijen. Prema tome pravilu, pojavnica može sadržavati bilo koje grafeme prije morfema *-ski*, ali mora sadržavati taj tvorbeni morfem i jedan od sljedećih sufikasa: *-ijima*, *-ijega*, *-ijemu*, *-ijem*, *-ijim*, *-ijih*, *-ijoj*, *-ijeg*, *-iji*, *-ije*, *-ija*, *-oga*, *-ome*, *-omu*, *-ima*, *-og*, *-om*, *-im*, *-ih*, *-oj*, *-i*, *-e*, *-o*, *-a* ili *-u*. S pomoću toga pravila razdvajamo oblike leksema kao što su *županijski* i *županija*, *županjski* i *Županja* itd. Treće navedeno pravilo trebalo bi svesti različite oblike glagola koji je tvoren tvorbenim sufiksom *-injati* na zajednički pseudokorijen. Prema ovome pravilu, pojavnica može sadržavati bilo koji niz prije morfema *inj*, ali mora sadržavati slijed grafema „inj" i jedan od sljedećih oblićnih nastavaka: *-asmo*, *-aste*, *-ati*, *-emo*, *-ete*, *-ali*, *-ala*, *-alo*, *-ali*, *-ale*, *-aše*, *-ahu*, *-em*, *-eš*, *-at*, *-ah* ili *-ao*.

4.3.1. Izrada pravila za imenice

Pravila koja su izrađena za imenice napravljena su na temelju popisa 527 imeničkih sufikasa iz *Tvorbe riječi u hrvatskome književnome jeziku* (Babić 2002: 70). Ta pravila možemo podijeliti u tri skupine: ona koja sadržavaju cijeli tvorbeni sufiks, ona koja sadržavaju skraćene ili proširene inačice tvorbenih sufikasa i ona koja obuhvaćaju nekoliko tvorbenih sufikasa. Pravila koja sadržavaju cijeli tvorbeni sufiks napravljena su s pomoću sljedećih sufikasa: *-aj(a)c*,

[10] Ovo pravilo obuhvaća i tvorbene sufikse *-ijan*, *-kan*, *-išan*, *-utan*, *-az(a)n*.

[11] Nulti morfem označen je znakom razmaka. Ako se na kraju pravila nalazi okomita crta, to znači da je u pravilo uključen i nulti morfem (*-ø*).

-jač, *-ič*, *-ać*, *-bać*, *-dać*, *-jad*, *-ljag*, *-ing*, *-juh*, *-ač(a)k*, *-arin*, *-ašin*, *-etin*, *-aist*, *-ov* i *-ez*. Već smo vidjeli kako izgleda pravilo napravljeno koristeći cijeli tvorbeni sufiks (.+ost ima|i|), ali ako tvorbeni sufiks završava samoglasnikom, moramo ga prilagoditi. Slično kao što osnovu imenice dobivamo tako da se od nje odijeli jedninski genitivni nastavak[12] (*vrapc-a*, *pism-a*, *izvedb-e*) (Barić i dr. 1995: 289), tako je prilagodbom sufiksa *-idba* nastalo pravilo:

.+idb ama|om|a|e|i|u|o

Desni dio toga pravila sadržava oblične nastavke kojima je obuhvaćena većina oblika imenica koje se tvore tvorbenim sufiksom *-idba*. Oblici *selidba*, *selidbe*, *selidbi*, *selidbu*, *selidbo* i *selidbom* trebali bi s pomoću ovoga pravila biti svedeni na pseudokorijen *selidb*. Jedino oblik *selidaba* nije obuhvaćen tim pravilom, ali taj smo nedostatak riješili s pomoću tranformacije „daba → dba”. [13] Osim tvorbenoga sufiksa *-idba*, sljedeći su sufiksi prilagođeni ukljanjanjem završnoga samoglasnika: *-čica*, *-alica*, *-elica*, *-ilica*, *-uljica*, *-nica*, *-arica*, *-urica*, *-ašica*, *-ušica*, *-ašica*, *-jetica*, *-otica*, *-ovca*, *-oća*, *-ijada*, *-erda*, *-čaga*, *-jaga*, *-jeha*, *-juha*, *-čija*, *-anija*, *-oja*, *-ačka*, *-aljka*, *-iljka -ojka*, *-enka*, *-eska*, *-itka*, *-uka*, *-alja*, *-ilja*, *-ulja*, *-bina*, *-čina*, *-šćina*, *-anja*, *-ura*, *-onja*, *-ijera*, *-ura*, *-esa*, *-isa*, *-eša*, *-ava*, *-itva*, *-drva*, *-eza*, *-ašce*, *-ešce*, *-enče*, *-ašće*, *-evlje* i *-ovlje*. Primjer takvoga pravila jest:

.+ovlj ima|em|a|e|u[14]

Pravila koja sažimaju nekoliko imeničkih tvorbenih sufikasa u jedan napravljena su uporabom sljedećih sufikasa: *-nica*, *-vica*, *-jača*, *-urda*, *-jaga*, *-uga*, *-cija*, *-dija*, *-lija*, *-sija*, *-jika*, *-anka*, *-inka*, *-arka*, *-uška*, *-elja*, *-vina*, *-dra*, *-ura*, *-uša*, *-java*, *-jač*, *-enje*, *-ište*, *-jaj*, *-ij*, *-jak*, *-nik*, *-elj*, *-ilo*, *-stvo*, *-ar*, *-or*, *-itis*, *-jaš*, *-oš*, *-et*, *-ut* i *-ež*. Primjer takvoga pravila jest:

.+ušk ama|om|a|e|i|u|o[15]

Sljedeći tvorbeni sufiksi svrstani su u skupine na temelju zajedničkih dijelova: *-ant*, *-ijant*, *-tant* i *-ent*; *-džija*, *-adžija* i *-edžija*; *-erija*, *-merija*, *-orija*, *-urija* i *-jurija*; *-sija*, *-esija* i *-osija*; *-atika* i *-etika*; *-avka* i *-ovka*; *-ačina*, *-jačina* i *-ičina*; *-alina* i *-olina*; *-arina*, *-erina*, *-urina*, *-čurina* i *-jurina*; *-esina* i *-usina*; *-je-*

[12] Osnove imenica u pravilu završavaju suglasnikom, a iznimke su imenice koje su došle u hrvatski iz drugih jezika. Osnove imenica kao što su *hobi*, *ragu* i *kakadu* završavaju samoglasnikom (Silić i Pranjković 2005: 104).

[13] Više o transformacijama reći ćemo u poglavlju 4.4.

[14] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *jezikoslovlje*.

[15] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *bjelouška*.

šina i *-ušina*; *-etina*, *-ičetina*, *-otina*, *-jotina*, *-ština*, *-inština*, *-uština*, *-ovština* i *-utina*; *-eskara* i *-uskara*; *-enjara* i *-onjara*; *-atika*, *-atik*, *-etika* i *-etik*; *-čanin*, *-ičanin*, *-ančanin*, *-čančanin*, *-jančanin*, *-ovljančanin*, *-inčanin*, *-arčanin*, *-jarčanin*, *-evčanin* i *-ovčanin*; *-ćanin*, *-ićanin* i *-šćanin*; *-janin*, *-ijanin*, *-ljanin*, *-eljanin*, *-evljanin*, *-ovljanin*, *-anjanin* i *-janjanin*; *-aranin*, *-jaranin*, *-oranin*; *-ašanin* i *-štanin*. Primjer takvoga pravila jest:

.+(b|c|e|g|j|k|s)an kama|cima|kom|aca|aka|cem|ac|ca|cu|če|ci|ce|ka|ke|ki|ku|ko

Sljedeći su tvorbeni sufiksi ili prošireni ili skraćeni u svrhu izrade pravila: *-uljica*, *-ija*, *-ija*$_2$, *-agija*, *-arija*, *-sija*, *-aja*, *-jaja*, *-nina*, *-arna*, *-ara*, *-jara*, *-aj*, *-eta*, *-vanin*, *-ežanin*, *-evo* i *-aštvo*. Primjer takvoga pravila jest:

.+tnin ama|om|a|e|u|o[16]

.+ket ama|om|a|u|e|i[17]

Dodana su pravila koja počinju s „.+njic", „.+arica", „.+žaj", „.+šaj", „.+etk", „.+kn", „.+enj", „.+aša", „.+ket", „.+tet", i „.+drac". Neka od njih izrađena su na temelju sufiksoida. Na temelju sufiksoida *-slovlje*, kojim se tvore leksemi kao što su *jezikoslovlje*, *prirodoslovlje* i *mudroslovlje*, nastalo je pravilo „.+ovlj ima|em|a|e|u". Na temelju sufiksoida *-logija*, kojim se tvore leksemi kao što su *arheologija*, *ekologija* i *ideologija*, nastalo je pravilo „.+gij ama|om|a|u|e|i|o". Iako je većina pravila oblikovana tako da traženi pseudokorijen bude jednak osnovi imenice, katkad su, u slučajevima kada dolazi do glasovnih promjena u osnovi riječi, bolje rezultate davala pravila čiji završetak nije bio jednak osnovi riječi. Primjer takvoga pravila jest:

.+itis cima|aka|kom|ku|ke|ak|ci[18]

U tablici 2 nalazi se popis sufikasa koji sudjeluju u tvorbi imenica. Debelo su označeni sufiksi, tj. njihovi dijelovi, koji su svrstani u jedinstveno pravilo korjenovanja kada završavaju istim slijedom grafema (npr. *-cija*, *-acija*, *-ijacija*, *-ancija* i *-encija* korišteni su za pravilo .+ij ima|ama|om|a|e|i|u|o|), a kosim su slovima označeni sufiksi, tj. njihovi dijelovi, koji su korišteni za izradu pojedinačnih pravila korjenovanja. Sufiksi koji nisu nisu korišteni za izradu pravila korjenovanja nisu označeni.

[16] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *nekretnina*.
[17] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *anketa*.
[18] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *pritisak*.

-a	-*ija*$_2$	-*ilja*	-at**ura**	-l**jač**	-enj(a)k	**-čan**in	-jan**čar**
-a$_2$	-**cija**	-*ulja*	-ijat**ura**	-*ič*	-ar(a)k	-i**čan**in	-**ćar**
-ba	-a**cija**	-ima	-*esa*	-oč	-er(a)k	-an**čan**in	-j**ar**
-*idba*	-ija**cija**	-na	-*isa*	-*ać*	-ur(a)k	-čan**čan**in	-l**ar**
-oba	-i**cija**	-*ana*	-ša	-*bać*	-čur(a)k	-jan**čan**in	-ion**ar**
-ca	-an**cija**	-ijana	-*eša*	-*dać*	-eš(a)k	-ovljan**ča-n**in	-t**ar**
-ica	-en**cija**	-ična	-iša	-ić	-uš(a)k	-in**čan**in	-uš**ar**
-čica	-*čija*	-ina	-**uša**	-**čić**	-at(a)k	-ar**čan**in	-er
-ajica	-a**dija**	-ina$_2$	-e**uša**	-i**čić**	-et(a)k	-jar**čan**in	-*ander*
-lica	-ur**dija**	-*bina*	-k**uša**	-ušić	-it(a)k	-ev**čan**in	-*ijer*
-*alica*	-u**dija**	-**čina**	-ta	-u*tić*	-ot(a)k	-ov**čan**in	-*ioner*
-*elica*	-džija	-a**čina**	-eta	-je**vić**	-ut(a)k	-**ćan**in	-*ater*
-*ilica*	-adžija	-ja**čina**	-*ota*	-o**vić**	-āk	-i**ćan**in	-*ezer*
-olica	-edžija	-i**čina**	-*ista*	-ad	-čāk	-š**ćan**in	-ir
-u*ljica*	-a*gija*	-*šćina*	-*osta*	-*jad*	-šćāk	-**jan**in	-**or**
-***nica***	-**lija**	-a**lina**	-va	-and	-išćāk	-i**jan**in	-t**or**
-a**nica**	-a**lija**$_1$	-o**lina**	-*ava*	-id	-**jāk**	-l**jan**in	-at**or**
-e**nica**	-a**lija**$_2$	-*nina*	-**java**	-e$_1$	-l**jāk**	-el**jan**in	-ikat**or**
-al**nica**	-i**lija**	-a**rina**	-n**java**	-e$_2$	-n**jāk**	-evl**jan**in	-it**or**
-il**nica**	-aj**lija**	-e**rina**	-*itva*	-ce	-in**jāk**	-ovl**jan**in	-*ur*
-o**nica**	-o**lija**	-u**rina**	-*drva*	-ice	-štāk	-an**jan**in	-**itis**
-ao**nica**	-jur**lija**	-ču**rina**	-*eza*	-ance	-ik	-jan**jan**in	-al**itis**
-io**nica**	-*anija*	-ju**rina**	-iza	-ence	-īk	-a**ran**in	-us
-ar**nica**	-a**rija**	-e**sina**	-oza	-*ašce*	-aik	-ja**ran**in	-aš
-ur**nica**	-e**rija**	-u**sina**	-aža	-*ešce*	-čik	-o**ran**in	-**jaš**
-uš**nica**	-me**rija**	-e**šina**	-uža	-če	-ščik	-a**šan**in	-*kaš*
-ov**nica**	-o**rija**	-je**šina**	-(a)c	-anče	-**nīk**	-š**tan**in	-io*naš*
-njica	-u**rija**	-u**šina**	-ič(a)c	-*enče*	-a**nīk**	-**van**in	-n**jaš**
-*arica*	-ju**rija**	-e**tina**	-j(a)c	-inče	-ja**nīk**	-**ežan**in	-iš
-*urica*	-***sija***	-iče**tina**	-*aj(a)c*	-će	-***enīk***	-*arin*	-**oš**$_1$
-*ašica*	-e**sija**	-o**tina**	-ej(a)c	-*ašće*	-be**nīk**	-*ašin*	-**oš**$_2$
-*ušica*	-o**sija**	-jo**tina**	-oj(a)c	-je	-il**nīk**	-*etin*	-ar**oš**
-*jetica*	-*oja*	-**ština**	-l(a)c	-le$_1$	-o**nīk**	-on	-uš
-*otica*	-ka	-inš**tina**	-al(a)c	-le$_2$	-ar**nīk**	-un	-at
-**vica**	-aka	-uš**tina**	-el(a)c	-*evlje*	-iš**nīk**	-(a)nj	-i*jat*

-a**vica**	*-ačka*	-ovš**tina**	-il(a)c	*-ovlje*	-av**nīk**	-o$_1$	-i*kat*
-o**vica**	-ika	-u**tina**	-elj(a)c	-nje	-ov**nīk**	-o$_2$	**-et**$_1$
-ovca	**-jika**	**-vina**	-an(a)c	**-enje**	-a**tik**	-co	**-et**$_2$
-ača	-l**jika**	-a**vina**	-čan(a)c	-j**enje**	-e**tik**	-jo	-ol**et**
-jača	-atika	-ja**vina**	-jan(a)c	**-inje**	-ok	-ojo	-it**et**
-ul**jača**	-etika	-lja**vina**	-šćan(a)c	-l**inje**	-luk	-ko	*-it*
-n**jača**	*-ojka*	-je**vina**	-išćan(a)c	**-ište**	-al	-ajko	-a**nt**
-aća	-alka	-o**vina**	-ijan(a)c	-l**ište**	-(a)lj	-lo	-ija**nt**
-oća	*-aljka*	-ajna	-ljan(a)c	-el**ište**	-alj	-alo	-ta**nt**
-da	*-iljka*	-ona	-itan(a)c	-il**ište**	**-elj**$_1$	**-ilo**	-e**nt**
-ada	**-anka**	*-arna*	-en(a)c	-ov**ište**	**-elj**$_2$	-ov**ilo**	-ot
-ijada	-č**anka**	-nja	-ijen(a)c	*-ljag*	-t**elj**	-mo	-st
-enda	-j**anka**	*-anja*	-in(a)c	*-ing*	-at**elj**	-no	-est
-ur**enda**	-št**anka**	**-inja**	-vin(a)c	-uh	-it**elj**	-ino	-ist
-erda	*-enka*	-k**inja**	-on(a)c	*-juh*	-ilj	-ro	*-aist*
-urda	**-inka**	-ot**inja**	-un(a)c	***-aj***	-ulj	-šo	**-ost**
-j**urda**	-al**inka**	*-onja*	-anj(a)c	**-jaj**	-iz(a)m	-eto	-k**ost**
-aga	**-arka**	*-ara*	-ar(a)c	-l**jaj**	-ioniz(a)m	**-ev**o	-n**ost**
-čaga	-j**arka**	*-jara*	-jar(a)c	**-ij**	-im	-ivo	**-ut**
-jaga	*-eska*	-e**skara**	-er(a)c	-ar**ij**	-an	-ovo	-e**ut**
-l**jaga**	**-uška**	-u**skara**	-or(a)c	-or**ij**	-ijan	**-stvo**	-av$_1$
-ul**jaga**	-j**uška**	-e**njara**	-uš(a)c	-(a)k	-kan	-an**stvo**	-av$_2$
-uga	*-itka*	-o**njara**	-ev(a)c	*-ač(a)k*	-išan	-in**stvo**	*-ov*
-č**uga**	*-uka*	**-dra**	-ov(a)c	-ečˇ(a)k	-utan	-a*štvo*	*-ez*
-j**uga**	-a**vka**	-en**dra**	-kov(a)c	-ičˇ(a)k	-az(a)n	-zo	**-ež**$_1$
-jeha	-o**vka**	*-ijera*	-ikov(a)c	-ad(a)k	-en$_1$	-up	**-ež**$_2$
-juha	-la	-ira	-ac	-j(a)k	-en$_2$	-r	-j**ež**
-ja	-lja	***-ura***	-ic	-elj(a)k	-in	**-ar**	-iž
-aja	*-alja*	-d**ura**	-ač	-ulj(a)k	-in	-č**ar**	*-ø*
-jaja	**-elja**	-j**ura**	-ač$_2$	-julj(a)k	-elin	-ič**ar**	-ø$_2$
-ija	-k**elja**	-t**ura**	***-jač***	-anj(a)k	-anin	-ničˇ**ar**	

Tablica 2. Popis tvorbenih sufikasa imenica

4.3.2. Izrada pravila za pridjeve

Pravila koja su izrađena za pridjeve napravljena su na temelju popisa 160 pridjevskih tvorbenih sufikasa iz *Tvorbe riječi u hrvatskome književnome jeziku* (Babić 2002: 381). Ta pravila možemo podijeliti u tri skupine: ona koja sadržavaju cijeli tvorbeni sufiks, ona koja sadržavaju skraćene ili proširene inačice tvorbenih sufikasa i ona koja obuhvaćaju nekoliko tvorbenih sufikasa. Pravila koja sadržavaju cijeli tvorbeni sufiks oblikovana su uporabom sljedećih sufikasa: *-stvenī*, *-ovljī*, *-evljī*, *-škī*, *-ārnī*, *-enjī*, *-ijel(a)n*, *-ar(a)n*, *-čat*, *-nat*, *-ični*, *-beni*, *-meni*, *-ač(a)n*, *-uoz(a)n*, *-jan*. Primjer takvoga pravila jest:

.+ben ijima|ijega|ijemu|ijeg|ijem|ijim|ijih|ijoj|ima|oga|iji|ije|ija|iju|im|ih|oj|og|e|a|u|i|o[19]

Pravila koja sadržavaju skraćene ili proširene inačice tvorbenih sufikasa oblikovana su uporabom sljedećih sufikasa: *-(a)n*, *-et(a)n* i *-ov*. Primjer takvoga pravila jest:

.+metn ijima|ijega|ijemu|ijeg|ijem|ijim|ijih|ijoj|ovima|ove|ovi|ova|ima|iji|ije|ija|iju|oga omu|ome|ima|og|om|ih|im|o|i|e|a[20]

Pravila koja obuhvaćaju nekoliko tvorbenih sufikasa oblikovana su na temelju sljedećih sufikasa: *-skī*, *-alnī*, *-uš(a)n*, *-et(a)n*, *-ast*, *-kav*, *-jav*, *-šav*, *-ev*, *-jiv*, *-iv* i *-č(a)n*. Primjer takvoga pravila jest:

.+(s|š)k ijima|ijega|ijemu|ijem|ijim|ijih|ijoj|ijeg|iji|ije|ija|oga|ome|omu|ima|og|om|im|ih|oj|i|e|o|a|u[21]

U tablici 3 nalazi se popis sufikasa koji sudjeluju u tvorbi pridjeva. Debelo su označeni sufiksi, tj. njihovi dijelovi, koji su svrstani u jedinstveno pravilo korjenovanja kada završavaju istim slijedom grafema, a kosim su slovima označeni sufiksi, tj. njihovi dijelovi, koji su korišteni za izradu pojedinačnih pravila korjenovanja. Sufiksi koji nisu korišteni za izradu pravila korjenovanja nisu označeni.

[19] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *službe**n***.
[20] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *razmeta**n***.
[21] Primjer leksema na koji se odnosi navedeno pravilo korjenovanja: *hrvatsk**i***.

-uščī	-ijev**skī**	*-enjī*	*-ijel(a)n*	-at	-juš**kast**	-lj**ev**	*-meni*
-ećī	-ljev**skī**	-injī	*-ar(a)n*	-cat	-ol**ast**	-evlj**ev**	-ec(a)n
-ećī	-ov**skī**	-šnjī	-š(a)n	-acat	-olj**ast**	-ovlj**ev**	-ic(a)n
-jī	*-škī*	-ašnīi	-aš(a)n	-čat	-ulj**ast**	**-iv**	*-ač(a)n*
-ijī	*-evljī*	-(a)k	-jaš(a)n	-ijat	-nj**ast**	-č**iv**	**-ič(a)n**
-ujī	*-ovljī*	-ec(a)k	-eš(a)n	-ljat	-inj**ast**	-ećiv	-at**ič(a)n**
-ackī	-nī	-ic(a)k	-iš(a)n	*-nat*	-or**ast**	**-jiv**	-ah(a)n
-ačkī	*-stvenī*	-a**č(a)k**	-uš(a)n	-ovat	-uš**ast**	-l**jiv**	*-uoz*(a)n
-ičkī	**-ālnī**	-e**č(a)k**	-juš(a)n	-it	-njik**av**	-al**jiv**	-an
-skī	-ij**ālnī**	-aš(a)k	-at(a)n	-cit	-u**kav**	-el**jiv**	*-jan*
-e**skī**	-ion**ālnī**	-āk	***-et(a)n***	-evit	-l**jav**	-ol**jiv**	-en
-ej**skī**	-u**ālnī**	-ok	-ov**et(a)n**	-ovit	-el**jav**	-ažl**jiv**	-ven
-ij**skī**	-iōnī	-n	-ovj**et(a)n**	-novit	-ul**jav**	-ežl**jiv**	-in
-uj**skī**	*-ārnī*	*-(a)n*	-av(a)n	**-ast**	-n**jav**	-n**jiv**	-av
-an**skī**	-evnī	-jah(a)n	-iv(a)n	-ič**ast**	-in**jav**	*-ov*	*-ičav*
-jan**skī**	-ivnī	-j(a)n	-ov(a)n	-j**ast**	-on**jav**	-ani	**-jav**
-en**skī**	-ovnī	-esk(a)n	-az(a)n	-k**ast**	-un**jav**	-jani	**-kav**
-in**skī**	-ioznī	**-al(a)n**	-oz(a)n	-ik**ast**	-išav	*-ični*	-i**kav**
-inj**skī**	-njī	-u**al(a)n**	-atoz(a)n	-uškast	-ušav	-eni	-ji**kav**
-ev**skī**	-anjī	-ion**al(a)n**	-ioz(a)n	-juš**kast**	**-ev**	*-beni*	**-ø**

Tablica 3. Popis tvorbenih sufikasa pridjeva

4.4. Transformacije

Glasovne promjene mogu otežati rad korjenovatelja izrađenih za morfološki složene jezike kao što je hrvatski. Upravo je jedna od glavnih kritika korjenovanja nemogućnost rješavanja problema koje uzrokuju glasovne promjene. Kako bismo umanjili njihov utjecaj, koristili smo se transformacijama koje, kao što njihov naziv daje naslutiti, pretvaraju oblik leksema na kojima je primijenjena neka glasovna promjena u oblik koji korjenovatelj može lakše povezati s ostalim oblicima toga leksema.[22] Konzervativan korjenovatelj koji je oblikovan s pomoću pravila korjenovanja teško može na isti pseudokorijen svesti sve oblike imenice *prijedlog* (*prijedlog*, *prijedloga*, *prijedlogu*, *prijedlože*,

[22] Popis transformacija dostupan je i u Tadić (1994).

prijedlogom, *prijedlozi*, *prijedloge* i *prijedlozima*) bez transformacija. Transformacijama „lozi → loga" i „lozima → loga" poboljšali smo postupak korjenovanja za sve imenice koje završavaju sufiksom *-log* tako što smo uklonili moguće poteškoće prouzročene sibilarizacijom. Transformacijama „njaca → njca" i „njac → njca" omogućili smo korjenovatelju da svede sve oblike leksema *Županjac* na pseudokorijen *županjc*. Spomenuli smo već pravilo za imenice „.+idb ama|om|a|e|i|u|o". To pravilo obuhvaća većinu oblika leksema koji se tvore imeničkom sufiksom *-idba*, ali ipak ne sve. Genitiv množine *selidaba* nije obuhvaćen ovim pravilom. Pojavnice kao što je *selidaba* sadržavaju diskontinuirane ili prekinute morfove, tj. morfove prekinute nekim drugim morfom ili materijalom (Marković 2012: 42). Taj smo nedostatak riješili tranformacijom „daba → dba".

Kako bismo bili sigurni da se odabrane transformacije neće primijeniti na pogrešne pojavnice, provjeravali smo s pomoću Hrvatskoga nacionalnog korpusa[23] (HNK) koje sve lekseme određena transformacija može obuhvatiti. Tako smo, na primjer, s pomoću upita „.+daba" doznali da se u Hrvatskome nacionalnom korpusu (HNK_v25) nalazi 2498 pojavnica koje završavaju grafemima „daba" (*odredaba*, *primjedaba*, *izvedaba*, *naredaba*, *priredaba* itd.). Sve su navedene pojavnice oblici leksema koji se tvore imeničkim sufiksom *-dba*. Na temelju te provjere možemo zaključiti da tu transformaciju možemo primijeniti bez straha da će se primijeniti na pogrešne pojavnice. Naravno, HNK ne sadržava sve moguće oblike leksema hrvatskoga jezika, ali je vrlo dobar pokazatelj učinkovitosti pravila. Transformacije koje se mogu primijeniti nisu brojne upravo zato što katkad mogu napraviti više štete nego koristi. Šnajder, Dalbelo Bašić i Tadić (2009: 34) kao primjer problematičnoga leksema navode imenicu *vojnik* i njezine oblike *vojnika*, *vojniku*, *vojniče*, *vojnikom*, *vojnici*, *vojnike* i *vojnicima*. Transformacijom „nicima → nik" možemo oblik *vojnicima* svesti na pseudokorijen *vojnik*, ali nije preporučljivo koristiti se transformacijama da bismo sveli oblike *vojnici* i *vojniče* na isti pseudokorijen. Kada bismo upotrijebili transformaciju „nicima → nik", mnogim bi imenicama koje završavaju sufiksom *-ica* bio netočno određen pseudokorijen (npr. sljedeći dativi jednine *sjednici*, *stepenici*, *rečenici*, *sapunici* itd.); kada bismo upotrijebili transformaciju „niče → nik", brojni bi glagolski (npr. prezent 3. l. mn. *graniče*) i imenski (npr. akuzativ množine *braniče*) oblici bili svedeni na pogrešni pseudokorijen.

[23] Više o HNK-u na http://www.hnk.ffzg.hr i u Tadić (1997).

5. Vrednovanje korjenovatelja

5.1. Opis postupka vrednovanja

Kako bismo vrednovali konzervativan korjenovatelj čiju smo izradu opisali u četvrtome poglavlju (k2), usporedili smo njegovu preciznost, odziv i F1-mjeru s istim vrijednostima korjenovatelja k1. Da bismo to mogli učiniti, bio nam je potreban uzorak za provjeru. Upotrijebili smo Jutarnji_9775, korpus od 9775 pojavnica koji je već vertikaliziran, tj. u kojemu su već određene leme i morfosintaktičke oznake.[24] Budući da je korpus Jutarnji_9775 vertikaliziran automatskim postupkom, provjerili smo jesu li ispravno određene leme svih 9775 pojavnica i ispravili one koje su bile netočno određene, radi točnijega vrednovanja korjenovatelja. Odredili smo i morfosintaktičke oznake za 646 pojavnica kojima nisu bili dodijeljeni potpuni podatci. Budući da su leme i morfosintaktičke oznake pojavnica u korpusu Jutarnji_9775 provjerene ručno, možemo pretpostaviti da je vrednovanje ovim putem bilo vjerodostojno. Primjere redova toga korpusa nakon korjenovanja možemo vidjeti u tablici 4.

pojavnica	lema	morfosintaktička oznaka	pseudokorijen
gradu	grad	Ncmsl	grad
stanovnika	stanovnik	Ncmpg	stanovnik
pješački	pješački	Afpmsay-n	pješačk
sjedište	sjedište	Ncnsa	sjedišt
vozači	vozač	Ncmpn	vozač
svjetski	svjetski	Afpmsny	svjetsk

Tablica 4. Primjeri redova korpusa Jutarnji_9775 nakon korjenovanja

5.2. Preciznost, odziv i F1-mjera

Preciznosti, odzive i F1-mjere izračunali smo koristeći se programom[25] za vrednovanje koji sadržava dva rječnika s umetnutim rječnicima[26], rječnik pseudokorijenā i rječnik lemā. Rječnik pseudokorijenā sastoji se od ključa, tj. pseudokorijenā (npr. *pregledavanj*) i vrijednosti, tj. umetnutih rječnika (npr.

[24] Za više informacija o automatskome određivanju morfosintaktičkih oznaka za hrvatski jezik v. Agić i Tadić (2006).

[25] Program za vrednovanje rezultata korjenovanja napisao je dr. sc. Nikola Ljubešić koristeći se programskim jezikom Python.

[26] Takvu vrstu rječnika zovemo rječnik rječnikā.

{‘pregledavanje#N’: 13}). Njegovi se umetnuti rječnici sastoje od ključeva, tj. lema s kojima ih je korjenovatelj spojio i vrijednosti, tj. frekvencija spajanja pseudokorijena i leme. Dani su primjeri iz toga rječnika:

pregledavanj {‘pregledavanje#N’: 13}
hrvatsk {‘Hrvatska#N’: 6, ‘hrvatski#A’: 7}

Iz ovih primjera možemo vidjeti da su pojavnice iz korpusa Jutarnji_9775 koje je korjenovatelj k2 sveo na zajednički pseudokorijen *pregledavanj*, njih 13, oblici leksema *pregledavanje* i da im je dodijeljena oznaka o vrsti riječi, N, tj. da su to oblici imenice. Također možemo vidjeti da je na pseudokorijen *hrvatsk* svedeno 13 pojavnica. Šest od njih 13 ima određenu lemu *Hrvatska*, imenicu, dok sedam njih ima određenu lemu *hrvatski*, pridjev. U svrhu računanja *preciznosti* (engl. *precision*) pretpostavljamo da je lema koja je najčešće povezana s određenim pseudokorijenom ispravno povezana[27]. Drugim riječima, ako je pseudokorijen *hrvatsk* šest puta povezan s lemom *Hrvatska*, a sedam puta s lemom *hrvatski*, pretpostavit ćemo da je ispravno povezan s lemom *hrvatski*. Za računanje preciznosti trebaju nam dvije brojke iz toga rječnika. Jedna je od njih zbroj najčešćih, tj. „ispravnih” sparivanja između pojedinih pseudokorijena i lema (m), a druga je zbroj svih sparivanja (s). Izračunat ćemo preciznost (p) za primjere navedene iz rječnika pseudokorijenā.

$$m = 13 + 1 + 7 + 7 + 3 = 31$$
$$s = 13 + 1 + 6 + 7 + 7 + 3 = 37$$
$$p = \frac{m}{s} = 0{,}838$$

Rječnik lemā sastoji se od ključa, tj. lemā, i vrijednosti, tj. umetnutih rječnika koji se sastoje od pseudokorijena koji su povezani s tom lemom i njihovih frekvencija. Dani su primjeri iz toga rječnika:

pregledavanje#N {‘pregledavanj’: 13}
računalo#N {‘računal’: 7, ‘račun’: 3}

U svrhu računanja *odziva* (engl. *recall*) pretpostavljamo da je pseudokorijen koji je češće povezan s lemom ispravno povezan i da su sva druga sparivanja rezultat pogreške. Drugim riječima, ako je lema *računalo* sedam puta povezana

[27] Problem homografije nije uzet u obzir.

s pseudokorijenom *računal*, a tri puta s lemom *račun*, pretpostavit ćemo da je ispravno povezana s pseudokorijenom *računal*. Za računanje odziva koristimo zbroj frekvencija svih ispravno dodijeljenih pseudokorijena (f) i zbroj svih sparivanja (s) koji smo već koristili za računanje preciznosti. Izračunat ćemo odziv (r) za navedene primjere:

s = 37
f = 13 + 1 + 6 + 7 + 7 = 34
r = = = 0,919

Nakon što smo izračunali preciznost i odziv, možemo, za navedene primjere, izračunati i F1-mjeru, tj. harmonijski prosjek preciznosti i odziva:

$$F1 = 2 * \frac{preciznost * odziv}{preciznost + odziv} = 2 * \frac{p * r}{p + r} = 2 * \frac{0{,}838 * 0{,}919}{0{,}838 + 0{,}919} = 2 * \frac{0{,}77}{1{,}752} = 0{,}879$$

5.3. Poredbena analiza korjenovateljā

Nakon što smo izračunali odziv, preciznost i F1-mjeru za korjenovatelj k1 i konzervativni korjenovatelj k2, učinkovitost tih korjenovatelja ispitali smo i s pomoću dodatnih uvjeta. Prvo smo usporedili njihovu učinkovitost koristeći transformacije i bez njih. Njihovu smo učinkovitost također ispitali ovisno o vrstama riječi na koje se primjenjuju – na sve vrste riječi, bez zaustavnih riječi[28] i samo na imenice i pridjeve. Od 9775 pojavnica u korpusu Jutarnji_9775 njih 1210 pravopisni su znakovi, njih 6028 su samoznačnice, a 2466 suznačnice. Na popisu zaustavnih riječi za hrvatski jezik obično se nalaze zamjenice, prijedlozi, veznici, brojevi, čestice, uzvici, pomoćni oblici glagolā *biti* i *htjeti* te modalni glagoli *željeti*, *morati*, *trebati* i *moći*. Nije bilo potrebno napraviti potpuni popis zaustavnih riječi s obzirom na to da su vrste riječi već označene u korpusu Jutarnji_9775 i da program za vrednovanje nudi mogućnost određivanja koje ćemo sve vrste riječi uključiti tijekom računanja preciznosti, odziva i F1-mjere.

[28] U obradi se prirodnoga jezika funkcionalne riječi nazivaju *zaustavne riječi*, prema engleskome nazivu *stop words*. Silić i Pranjković (2005) u *Gramatici hrvatskoga jezika* za funkcionalne riječi predlažu naziv *suznačnice*. Oni riječi koje imaju samostalno značenje definiraju kao samoznačne riječi ili samoznačnice, a riječi koje nemaju samostalno značenje kao suznačne riječi ili suznačnice.

Istom smo metodom računali učinkovitost korjenovatelja kada se uzimaju u obzir samo imenice i pridjevi. Najbolje smo rezultate dobili kada smo ograničili korjenovanje na imenice i pridjeve, dvije najvažnije vrste riječi za pretraživanje informacija. Prikaz tih rezultata možemo vidjeti u dijagramu 1. U tablici 6 možemo usporediti preciznosti, odzive i F-1 mjere korjenovateljā k1, k2 i k3.

6. Optimizacija korjenovatelja

6.1. Izrada korjenovatelja k3

Budući da rezultati vrednovanja korjenovatelja k2 nisu bili zadovoljavajući, odlučili smo pojednostavniti pravila i smanjiti njihov broj. Naša je pretpostavka bila da se s manjim brojem pravila može jednostavnije manipulirati i time lakše doći do pravila koji daju najbolje rezultate. Od 301 pravila izbacili smo ona koja nemaju znatan utjecaj na rezultate, a mnoga smo i spojili. Tako su npr. četiri pravila: „.+(ž|š)aj ima|em|a|e|i|u| ”, „.+učaj evima|evi|eva|eve|em|a|u| ”, „.+jaj ima|om|a|u|e|i| ” i „.+ostaj ama|om|a|u|e|i” pretvorena u pravilo „.+(t|č|j|ž|š)aj ev ima|evi|eva|eve|ama|ima|em|a|e|i|u| ”. Odbačena su pravila za lekseme čiji se oblici vrlo rijetko pojavljuju u Hrvatskome nacionalnom korpusu, npr. „.+naut ima|om|a|u|e|i| ” i „.+ijazm ima|om|i|a|u|e”. Ispitivanjem učinkovitosti pojedinih pravila, tj. njihovim privremenim isključivanjem iz postupka korjenovanja, došli smo do zaključka da je većina proširenih pravila za imenice nepotrebna dok god one ne završavaju grafemima kojima završavaju pravila oblikovana za pridjeve ili glagole. Istom smo metodom također došli do zaključka da se najsloženija pravila isplati oblikovati za glagole i pridjeve. Kao i za prošle inačice korjenovatelja, za provjeru i vrednovanje pravila korišten je korpus Jutarnji_9775. Da bismo izbjegli oblikovanje pravila koja dobivaju dobre rezultate samo na tome korpusu, ista smo pravila provjeravali s pomoću korpusa Index_1000. Na taj smo način dobili učinkovitiji korjenovatelj sa znatno manjim brojem pravila, njih 73, od kojih je 35 za imenice, 14 za pridjeve i 24 za glagole.

Smanjen je i broj transformacije sa 140 na 123. Jedan primjer pojednostavnjenja jest zamjena transformacija „bošću → bosti”, „došću → dosti”, „tošću → tosti”, „jošću → josti”, „košću → kosti”, „lošću → losti”, „nošću → nosti”, „rošću → rosti” i „vošću → vosti” transformacijom „ošću → osti”. Prethodno je prednost bila dana većemu broju transformacija zato što ona izbjegavaju pogreške poput pretvaranja oblika „gošću”, „Živogošću” i „Vogošću” u s njima nepovezane oblike „gosti”, „Živogosti” i „Vogosti”.

6.2. Izrada pravila za glagole

Za novu inačicu korjenovatelja izrađena su detaljnija pravila za glagole. Iako glagoli nemaju istu važnost za pretraživanje informacija kao imenice i pridjevi, naša je pretpostavka da detaljnija pravila za korjenovanje glagole mogu pridonijeti točnijemu korjenovanju drugih vrsta riječi. Koristili smo se tvorbenim nastavcima koje Babić navodi u *Tvorbi riječi u hrvatskome književnome jeziku* (2002: 503). Najuspješnija su pravila izrađena za petu vrstu glagola[29], tj. glagole kojima infinitivna osnova sadržava morf *a* (infinitiv na *-ati*), osim onih koji u prezentu imaju nastavke s morfom *i*. Uspješno su oblikovana pravila za ostale vrste glagola, osim za nepravilne glagole kao što su *ići* i *izaći*.[30] U tablici 5 nalaze se svi tvorbeni sufiksi korišteni za izrade pravila za glagole. Debelo su označeni sufiksi, tj. njihovi dijelovi, koji su svrstani u jedinstveno pravilo, a kosim su slovima označeni sufiksi, tj. njihovi dijelovi, koji su korišteni, ali nisu svrstavani zajedno s drugim sufiksima u pojedinačno pravilo.

-ati	-ka**rati**	-atati	-o**vati**
-cati	**-irati**	**-etati**	-ko**vati**
-u**cati**	-fici**rati**	-k**etati**	-iko**vati**
-čati	-ifici**rati**	*-otati*	-jeti
-u*dati*	-iz**irati**	*-utati*	-iti
-jati	**-sati**	-vati	-čiti
-kati	-a**sati**	**-avati**	-ačiti
-a**kati**	-e**sati**	-j**avati**	*-ičiti*
-ja**kati**	-i**sati**	**-evati** (-ujem)	*-uljiti*
-uc**kati**	-adi**sati**	**-evati** (-evam)	**-ariti**
-i**kati**	-o**sati**	-ij**evati**	-č**ariti**
-u*škati*	-u**sati**	**-ivati** (-ivam)	-k**ariti**
-ukati	**-šati**	**-ivati** (-ujem)	-u*šiti*
-ij**ukati**	-u**šati**	-j**ivati**	**-nuti**
-in*jati*	-tati	-k**ivati**	-u**nuti**
-a**rati**			

Tablica 5. Popis glagolskih tvorbenih sufikasa

[29] Težak i Babić (2005), Barić i dr. (2005), Raguž (1997) te Silić i Pranjković (2005) različito dijele vrste glagola. U ovome se smo radu koristili podjelom *Hrvatske gramatike* Eugenije Barić i suradnika.

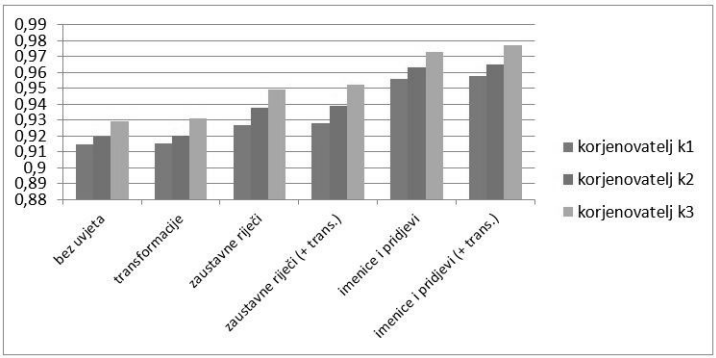
[30] Rješenja za nepravilne glagole dostupna su u Tadić (1994).

6.3. Vrednovanje korjenovatelja k3

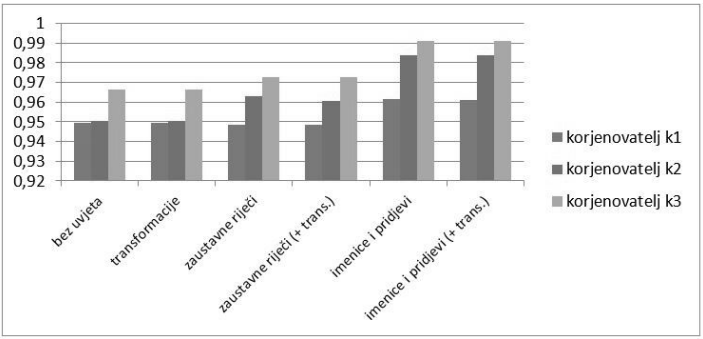
Za vrednovanje nove inačice korjenovatelja, tj. za izračun preciznosti, odziva i F1-mjere, ponovno smo se koristili programom za vrednovanje. Izračunali smo navedene vrijednosti nekoliko puta koristeći uvjete kao što su transformacije, uključenost zaustavnih riječi i vrste riječi na koje se izračun odnosi. Kao i korjenovatelji k1 i k2, korjenovatelj k3 postiže najbolje rezultate kada se izračun, uz korištenje transformacija, ograniči na imenice i pridjeve. U tablici 6 možemo vidjeti da korjenovatelj k3, s F1-mjerom od 0,977, postiže bolje rezultate od početnoga korjenovatelja k1 i od korjenovatelja k2, naprama F1-mjeri 0,957 korjenovatelja k1 i F-1 mjeri 0,965 korjenovatelja k2, tijekom korjenovanja imenica i pridjeva s uključenim transformacijama:

korjenovatelj	preciznost	odziv	F1-mjera
k1 (bez dodatnih uvjeta)	0,949	0,882	0,914
k2 (bez dodatnih uvjeta)	0,950	0,890	0,919
k3 (bez dodatnih uvjeta)	0,966	0,895	0,929
k1 s transformacijama	0,949	0,883	0,915
k2 s transformacijama	0,950	0,892	0,920
k3 s transformacijama	0,967	0,898	0,931
k1 bez transformacija i zaustavnih riječi	0,948	0,906	0,927
k2 bez transformacija i zaustavnih riječi	0,963	0,914	0,938
k3 bez transformacija i zaustavnih riječi	0,972	0,927	0,949
k1 s transformacijama, bez zaustavnih riječi	0,948	0,909	0,928
k2 s transformacijama, bez zaustavnih riječi	0,960	0,918	0,939
k3 s transformacijama, bez zaustavnih riječi	0,973	0,932	0,952
k1 bez transformacija (samo imenice i pridjevi)	0,962	0,950	0,956
k2 bez transformacija (samo imenice i pridjevi)	0,984	0,943	0,963
k3 bez transformacija (samo imenice i pridjevi)	0,991	0,955	0,973
k1 s transformacijama (samo imenice i pridjevi)	0,961	0,954	0,957
k2 s transformacijama (samo imenice i pridjevi)	0,984	0,947	0,965
k3 s transformacijama (samo imenice i pridjevi)	0,991	0,964	0,977

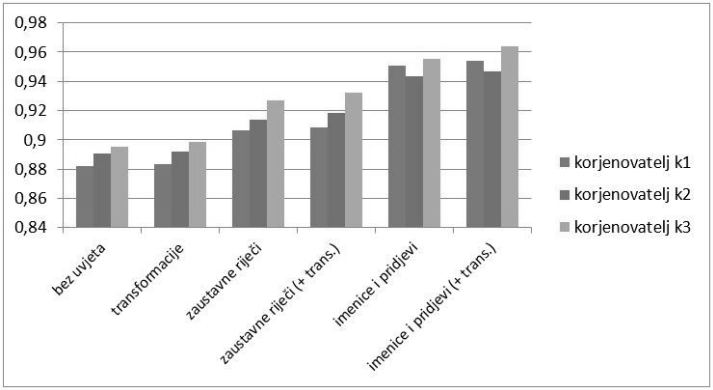
Tablica 6. Preciznosti, odzivi i F-1 mjere korjenovateljā k1, k2 i k3



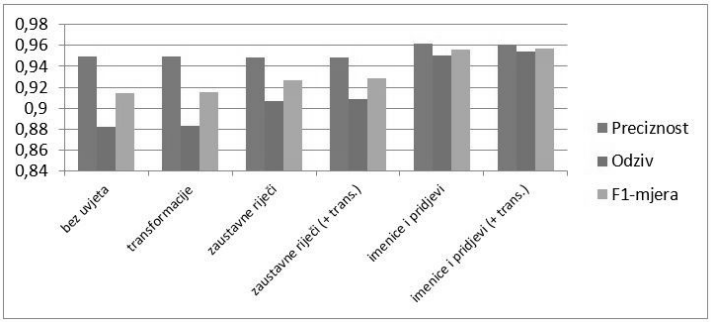
Dijagram 1. Izračun F1-mjere uz različite uvjete



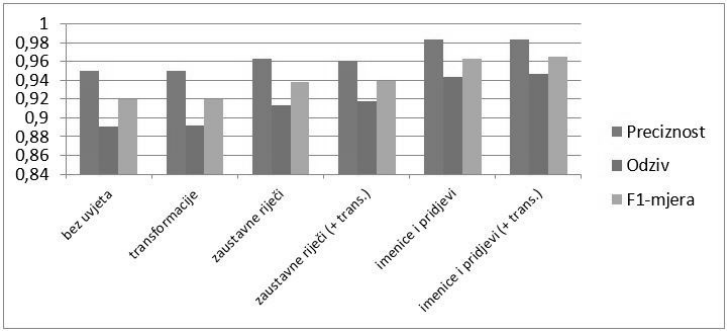
Dijagram 2. Izračun preciznosti uz različite uvjete



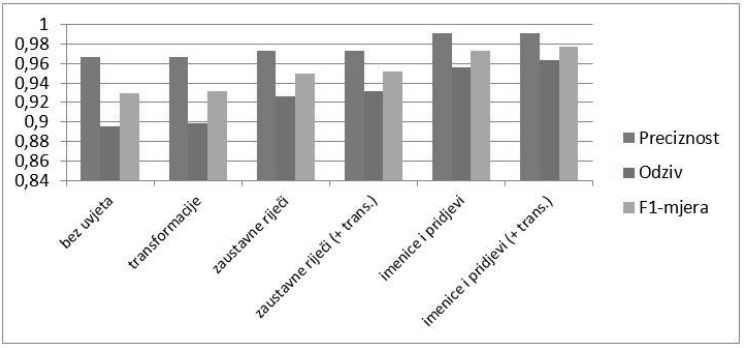
Dijagram 3. Izračun odziva uz različite uvjete



Dijagram 4. Preciznost, odziv i F-mjera za korjenovatelj k1



Dijagram 5. Preciznost, odziv i F-mjera za korjenovatelj k2



Dijagram 6. Preciznost, odziv i F-mjera za korjenovatelj k3

7. Analiza pogrešaka

S pomoću programa za vrednovanje analizirali smo pogreške korjenovateljā, tj. precizno utvrdili u kojim slučajevima korjenovatelj k1 i korjenovatelj k3 netočno određuju pseudokorijene pojavnica. Pogreške korjenovanja korjenovateljem k1 na korpusu Jutarnji_9775 prikazujemo u obliku rječnika rječnikā:

```
{‘A’: {‘A’: 6, ‘R’: 20, ‘V’: 9, ‘N’: 22}, ‘Y’: {‘Y’:
2}, ‘V’: {‘A’: 7, ‘V’: 2, ‘R’: 2, ‘N’: 12}, ‘R’: {‘A’:
18, ‘R’: 3, ‘V’: 3, ‘N’: 3}, ‘N’: {‘A’: 45, ‘Y’: 1,
‘R’: 7, ‘N’: 13, ‘V’: 19}}
```

Ako znamo da je ključ A oznaka za pridjev, ključ R oznaka za prilog, ključ V oznaka za glagol, ključ Y oznaka za kratice i ključ N oznaka za imenice, iz vrijednosti ključa A, tj. „{‘A’: 6, ‘R’: 20, ‘V’: 9, ‘N’: 22})”, možemo iščitati da je korjenovatelj k1 za šest pridjeva, 20 priloga, devet glagola i 22 imenice pogrešno odredio pseudokorijen, i da je taj pseudokorijen bio pseudokorijen pridjeva. U svrhu analize pogrešaka, pretpostavljamo da je pseudokorijenu kojemu je određena veza s više od jedne vrste riječi pravilno određena veza s onom vrstom riječi s kojom je najčešće doveden u vezu i da su ostala sparivanja rezultat pogreške. Ističe se velik broj pogrešaka korjenovatelja k1 u određivanju pseudokorijena imenicama koje pripadaju pridjevima.

Pogreške korjenovanja korjenovateljem k3 na korpusu Jutarnji_9775 prikazujemo u obliku rječnika rječnikā:

```
{‘A’: {‘A’: 6, ‘N’: 2, ‘R’: 20, ‘V’: 6}, ‘Y’: {‘Y’: 2},
‘V’: {‘A’: 5, ‘V’: 3, ‘R’: 3, ‘N’: 8}, ‘R’: {‘A’: 18,
‘V’: 1, ‘R’: 2, ‘N’: 2}, ‘N’: {‘A’: 4, ‘Y’: 2, ‘R’: 5,
‘V’: 10, ‘N’: 5}}
```

Usporedbom rezultata možemo uočiti da je u rezultatima korjenovatelja k3 manjemu broju imenica i glagola pogrešno određen pseudokorijen nekoga pridjeva, da je manjemu broju imenica pogrešno određen pseudokorijen nekoga glagola i da je manjemu broju glagola, imenica i pridjeva pogrešno određen pseudokorijen neke (druge) imenice. Najznačajnija promjena dogodila se u broju pridjeva kojima je pogrešno određen pseudokorijen neke imenice. Ipak, taj rezultat nije veliko iznenađenje s obzirom na to da je k1 agresivan korjenovatelj, a k3 konzervativan korjenovatelj. Korjenovatelj k3 s pomoću pravila „.+n ijima|ijega|ijemu|ijeg|ijem|ijim|ijih|ijoj|iji|ije|ija|iju|ima|ome|omu|oga|oj|om|ih|im|og|a|

e|i|o|u" i „.+sk ijima|ijega|ijemu|ijeg|ijem|ijim|ijih|ijoj|iji|ije|ija|iju|ima|ome|omu |oga|oj|om|ih|im|og|a|e|i|o|u" preciznije određuje pseudokorijene pridjeva.

pojavnica	korjenovatelj k1	korjenovatelj k3
medijskih	medij	medijsk
veličanstvenih	veličanstv	veličanstven
službenih	služb	služben
mrežnu	mrež	mrežn
brojnim	broj	brojn

Tablica 7. Pseudokorijeni pridjeva nakon korjenovanja (k1 i k3)

Korjenovatelj k3 najčešće griješi pri korjenovanju pridjeva i priloga (20 + 18 = 38). Takve je pogreške teže spriječiti zato što prilozi i pridjevi često dijele oblike, npr. prilog *brzo* i nominativ, akuzativ i vokativ jednine pridjeva *brz* u srednjemu rodu, prilog *sporo* i NAV jd. sr. r. pridjeva *sporo*, prilog *suprotno* i NAV jd. sr. r. pridjeva *suprotan*, prilog *službeno* i NAV jd. sr. r. pridjeva *služben* itd. Pogreške između glagola i pridjeva mogli bismo smanjiti kada bismo razdvojili glagolske pridjeve trpne od glagola. Oni često prelaze i u službu pravih pridjeva, a budući da su u korpusu Jutarnji_9775 u pravilu označeni kao glagoli, to dovodi do određenoga broja pogrešaka. Moguće je da bismo broj pogrešaka također mogli smanjiti kada bismo uspješno, s pomoću pravila, spojili superlative pridjeva s njihovim odgovarajućim pozitivima.

8. Nazivlje korjenovanja

Uočljiva je nedosljednost nazivlja u hrvatskim gramatikama u području morfologije. Kuna (2006: 166), na primjer, navodi „tvorku", „formant", „osnovni nastavak", „tvorbeni nastavak", „završetak", „dometak", „dočetak" i „sufiksalni morfem" kao nazive koje su razni autori predlagali kao alternativu za *sufiks.* U morfološkoj analizi riječi dijelimo na *osnovu* i *nastavak*. Težak i Babić (2005: 90) osnovu su definirali kao „onaj dio riječi koji se u morfološkim promjenama ne mijenja, osim u određenim okolnostima kada se u završnici osnove proizvodi kakva glasovna promjena...", a nastavak kao „glas ili glasovni skup koji se dodaje osnovi za tvorbu gramatičkih oblika riječi". Barić i dr. (1995: 289) razlikuju *obličnu osnovu* i *tvorbenu osnovu*. Prema toj gramatici, oblična osnova jedne riječi postaje tvorbena osnova druge riječi kada se uključi u tvorbeni proces. Marković (2012: 53) predlaže korištenje naziva *baza* koji bi, prema njemu, bio koristan kao

krovni izraz za „obličnu osnovu i tvorbenu osnovu". Marković u istoj knjizi korijen definira kao „obavezni, neizostavni dio oblika riječi, onaj koji ostaje kad se odstrane svi ostali morfovi, onaj koji je zajednički sastavni element srodnih riječi, onaj koji je nositelj temeljnoga značenja i temeljnoga izraza leksema" (2012: 50). Prema tim definicijama, osnova genitiva jednine imenice *podskup* (*podskupa*) bio bi *podskup*, a korijen istoga oblika te imenice bio bi *skup*. Iako korjenovatelji zapravo u većini slučajeva traže osnovu, a ne korijen riječi, i dalje upotrebljavamo taj naziv zato što je ustaljen. Ovdje predlažemo nazive *uosnovitelj* i *osnovatelj* kao zamjenske za naziv *korjenovatelj* i naziv *pseudoosnova* kao zamjenski za naziv *pseudokorijen*. Iako naziv *korjenovatelj* nije precizan, u ovome smo članku ipak upotrebljavali taj naziv s obzirom na to da je već prihvaćen među stručnjacima koji se bave računalnom lingvistikom i obradom prirodnoga jezika.

9. Zaključak

Rezultati vrednovanja, tj. izračun preciznosti, odziva i F1-mjere, konzervativnoga korjenovatelja k2 pokazali su da on daje bolje rezultate od korjenovatelja k1, ali ne dovoljno dobre s obzirom na broj pravila koji upotrebljava. Transformacije nisu znatno utjecale na rezultate. Budući da rezultati nisu bili zadovoljavajući, odlučili smo pojednostavniti i usavršiti pravila. Nova inačica konzervativnoga korjenovatelja (k3) sa smanjenim brojem dorađenih pravila postigla je bolje rezultate. Usporedbom preciznosti, odziva i F1-mjere korjenovatelja k1, k2 i k3, s različitima uvjetima kao što su transformacije, uključenost zaustavnih riječi i vrste riječi na koje se izračun odnosi, dokazali smo da korjenovatelj k3 dosljedno daje bolje rezultate. Kao i korjenovatelji k1 i k2, korjenovatelj k3 postiže najbolje rezultate kada se izračun, uz korištenje transformacija, ograniči na imenice i pridjeve. Analizirali smo pogreške korjenovateljā k1 i k3 te zaključili da je prednost korjenovatelja k3 što njegova pravila učinkovitije razdvajaju različite vrste riječi i u velikoj mjeri smanjuju broj interferencija koje mogu nastati tijekom postupka korjenovanja.

Tijekom usavršavanja pravila korjenovanja došli smo do zaključka da su proširena pravila, tj. pravila koja upotrebljavaju tvorbene nastavke, učinkovita za rješavanje problematičnih slučajeva pridjeva i glagola, ali da nema potrebe da većina tvorbenih nastavaka za imenice bude uključena u proširena pravila. Potrebna su daljnja istraživanja kako bi se ovaj korjenovatelj usavršio, iako je već sada moguća njegova primjena[31]. Budući da je korjenovatelj k3 dorađen

[31] Program koji je rezultat ovoga istraživanja objavljen je na sljedećoj mrežnoj adresi: http://nlp.ffzg.hr/resources/tools/stemmer-for-croatian.

koristeći testni uzorak, trebalo bi ga vrednovati koristeći dodatni uzorak. Vjerujemo da će se područje morfološke normalizacije za hrvatski jezik nastaviti istraživati i da će se upotreba ovakvih programa i dalje širiti.[32]

Literatura:

Agić, Željko; Tadić, Marko. 2006. *Evaluating Morphosyntactic Tagging of Croatian Texts*. ELRA. Ženeva – Pariz.

Babić, Stjepan. 2002. *Tvorba riječi u hrvatskome književnome jeziku*. HAZU – Nakladni zavod Globus. Zagreb.

Barić, Eugenija i dr. 1995. *Hrvatska gramatika*. Školska knjiga. Zagreb.

Kolaković, Zrinka. 2007. Zastupljenost padeža u hrvatskome jeziku u pisanim i govornim tekstovima. *Lahor: časopis za hrvatski kao materinski, drugi i strani jezik* 4. 242–270.

Kuna, Branko. 2006. Nazivlje u tvorbi riječi. *Filologija* 46–47. 165–182.

Lauc, Davor; Lauc, Tomislava; Boras, Damir; Ristov, Strahil. 1998. Developing text retrieval system using robust morphological parsing. *Proceedings of 20th International Conference on Information Technology Interfaces, ITI '98*. Ur. Kalpić, Damir; Hljuz-Dobrić, Vesna. SRCE. Zagreb. 61–65.

Ljubešić, Nikola i dr. 2007. Retrieving Information in Croatian: building a simple and efficient rule-based stemmer. *Digital information and heritage*. Ur. Seljan, Sanja; Stančić, Hrvoje. Odsjek za informacijske znanosti Filozofskoga fakulteta Sveučilišta u Zagrebu. Zagreb. 313–320.

Marković, Ivan. 2012. *Uvod u jezičnu morfologiju*. Disput. Zagreb.

Raguž, Dragutin. 1997. *Praktična hrvatska gramatika*. Medicinska naklada. Zagreb.

Silić, Josip; Pranjković, Ivo. 2005. *Gramatika hrvatskoga jezika*. Školska knjiga. Zagreb.

Šarić, Frane i dr. 2005. Enhanced thesaurus terms extraction for document indexing. *Proceedings of the 27th International Conference on Information Technology Interfaces*. Ur. Dobrić, Vesna. Sveučilište u Zagrebu. Zagreb. 227–232.

Šnajder, Jan; Dalbelo Bašić, Bojana; Tadić, Marko. 2009. Lexical-based Morphological Normalisation and its Application to Croatian. *Technologies for the Processing and Retrieval of Semi-Structured Documents*. Ur. Tadić, Marko; Dalbelo Bašić, Bojana; Moens, Marie-Francine. Hrvatsko društvo za jezične tehnologije. Zagreb. 23–80.

[32] Zahvaljujem recenzentima na izrazito konstruktivnim i korisnim prijedlozima za poboljšanje rada. Ovaj rad financirala je Hrvatska zaklada za znanost projektom *Repozitorij metafora hrvatskoga jezika* (3624).

Šnajder, Jan; Dalbelo Bašić, Bojana. 2009. String Distance-Based Stemming of the Highly Inflected Croatian Language. *Proceedings of Recent Advances in Natural Language Processing (RANLP2009)*. Ur. Angelova, Galia i dr. Incoma. Shoumen. 411–415.

Šnajder, Jan. 2010. *Morfološka normalizacija tekstova na hrvatskome jeziku za dubinsku analizu i pretraživanje informacija*. Doktorska disertacija. Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu. Zagreb. 184 str.

Tadić, Marko. 1994. *Računalna obrada morfologije hrvatskoga književnog jezika.* Doktorska disertacija. Filozofski fakultet Sveučilišta u Zagrebu. Zagreb. 160 str.

Tadić, Marko. 1997. Računalna obrada hrvatskih korpusa: povijest, stanje i perspektiva. *Suvremena lingvistika* 43–44. 387–394.

Težak, Stjepko; Babić, Stjepan. 2005. *Gramatika hrvatskoga jezika*. Školska knjiga. Zagreb.

Hrvatski morfološki leksikon. 2005. Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. http://hml.ffzg.hr/hml/. (pristupljeno 10. rujna 2015.).

Hrvatski nacionalni korpus. 2005. Zavod za lingvistiku Filozofskoga fakulteta Sveučilišta u Zagrebu. http://www.hnk.ffzg.hr/pretraga.html. (pristupljeno 10. rujna 2015.).

Building a Croatian language stemmer

Abstract

The paper presents two conservative Croatian language stemmers, k2 and k3. These stemmers are based on the k1 stemmer, an aggressive Croatian language stemmer presented by Nikola Ljubešić in a 2007 paper. By introducing an expanded set of rules that use derivational morphemes of nouns, verbs, and adjectives to determine the stems of words, we hoped to create a more efficient stemmer. In order to test whether the k2 and k3 stemmers were more efficient than the k1 stemmer, we calculated their precision, recall, and F1-score using a 9775 token corpus, and compared the results with the precision, recall, and F1-score of the k1 stemmer.

Ključne riječi: korjenovanje temeljeno na pravilima, računalna lingvistika, obrada prirodnoga jezika, hrvatski jezik

Keywords: rule-based stemming, computational linguistics, natural language processing, Croatian language