Analiza leksičke baze usluge strojnog provjernika pravopisa

Kordić, Ivan

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1321

ANALIZA LEKSIČKE BAZE USLUGE STROJNOG PROVJERNIKA PRAVOPISA

Ivan Kordić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 1321

ANALIZA LEKSIČKE BAZE USLUGE STROJNOG PROVJERNIKA PRAVOPISA

Ivan Kordić

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

ZAVRŠNI ZADATAK br. 1321

Pristupnik:	**Ivan Kordić (0036539994)**
Studij:	Elektrotehnika i informacijska tehnologija i Računarstvo
Modul:	Računarstvo
Mentor:	prof. dr. sc. Gordan Gledec

Zadatak: **Analiza leksičke baze usluge strojnog provjernika pravopisa**

Opis zadatka:

Usluga strojne provjere pravopisa "ispravi.me - hrvatski akademski spelling checker" kontinuirano se unapređuje funkcionalnostima koje poboljšavaju iskustvo upotrebe. Usluga počiva na rječničkoj bazi i n-gramskom sustavu stvorenom na temelju tekstova prispjelih na obradu. Rječnička baza sastavljena je od općejezičnog i posebnojezičnog fonda. Cilj je završnog rada proučiti dosadašnji razvoj usluge strojne provjere pravopisa te definirati metodologiju i analizirati rječničku bazu u smislu povezanosti morfološki sličnih različnica, koristeći javno dostupne servise. Radu treba priložiti izvorni i izvršni kod razvijenog sustava te potrebnu dokumentaciju.

Rok za predaju rada: 14. lipnja 2024.

Sadržaj

Uvod ... 1

1. Opis usluge ... 2

2. Osnove gramatike ... 5

2.1. Red riječi u rečenici ... 6

2.2. Funkcionalni stilovi hrvatskoga standardnog jezika ... 7

3. Opis zadatka i cilja rada ... 9

4. Označavanje gramatičkih kategorija ... 10

5. Pristup i metodologija dodavanja gramatičkih kategorija riječima ... 13

6. Pokrivanje jezične baze ... 19

6.1. Prefiksacija ... 20

6.2. Sufiksacija ... 23

7. Programska analiza rečenice ... 27

Zaključak ... 31

Literatura ... 32

Sažetak ... 35

Summary ... 36

Uvod

Usluga strojne provjere pravopisa hrvatskoga jezika *ispravi.me* popularna je i javno dostupna usluga koju je koristilo, primjerice 2023., preko 600 tisuća korisnika za obradu preko sedam milijuna tekstova s ukupno 1,74 milijardi riječi [1].

U ovome radu analizira se leksička baza usluge *ispravi.me* s ciljem pronalaska mogućnosti poboljšanja usluge u pogledu gramatičke analize teksta. Usluga *ispravi.me* odlična je za uočavanje pravopisnih grešaka i zatipaka, a potencijal za nadogradnju nalazi se u široj analizi teksta, odnosno ne samo pravopisnoj analizi nego i gramatičkoj analizi.

Gramatička analiza teksta je analiza koja se vrši na temelju gramatičkih svojstava i pravila jezika. Kada čovjek analizira rečenicu, mora znati gramatičke kategorije riječi da bi donio sud o ispravnosti. Također, kada bi stroj gramatički analizirao rečenicu morao bi biti programiran na temelju jezičnih pravila.

U ovome radu se analiziraju trenutne funkcionalnosti usluge *ispravi.me*. Potom je sadržan presjek osnovnih gramatičkih i stilističkih pravila koja bi se mogla koristiti u strojnoj analizi rečenica. Kako bi stroj znao koristiti pravila jezika potrebno je da su riječima u tekstu pridodane gramatičke kategorije. Stoga je u radu opisano nekoliko mogućih pristupa u označavanju gramatičkih kategorija riječima (engl. *Part of Speech Tagging*) poput skrivenih Markovljevih lanaca i uvjetnih slučajnih polja koji koriste vjerojatnosti izračunate preko učestalosti pojavljivanja i konteksta riječi u rečenici. Leksička baza usluge *ispravi.me* mora sadržavati informacije o gramatičkim kategorijama riječi, a trenutno to nema. Zato je veliki dio rada korištenje javno dostupne baze *hrLex*, leksikona hrvatskih riječi s označenim gramatičkim kategorijama i učestalosti njihova pojavljivanja, za pokrivanje leksičke baze usluge *ispravi.me*. Pokrivanje riječi leksičke baze usluge *ispravi.me* koristi svojstva hrvatskoga jezika za pokrivanje riječi kojih nema u *hrLexu*. Na kraju, napravljena je jednostavna analiza teksta po pristupu učestalosti pojavljivanja riječi te se razmatra kojim smjerom poći dalje u svrsi poboljšavanja usluge *ispravi.me*.

1. Opis usluge

Strojna provjera pravopisa usluge *ispravi.me* primarno se temelji leksičkoj bazi s tri fonda riječi [2]:

- hrvatski općejezični fond
- hrvatski posebnojezični fond (imena, nazivi)
- engleski općejezični fond

Kontekstna provjera temelji se na n-gramima, najpopularnijem statističkom jezičnom modelu [3]. U takvom se sustavu rečenice dijele na n-grame. N-grami su nizovi od n simbola odnosno, u slučaju ove usluge, n riječi. Ako se rečenica „Ovo je primjer rečenice.“ dijeli na n-grame, dobivaju se četiri unigrama (*ovo*, *je*, *primjer*, *rečenice*), tri bigrama (*ovo je*, *je primjer*, *primjer rečenice*), dva trigrama (*ovo je primjer*, *je primjer rečenice*) i jedan tetragram (*ovo je primjer rečenice*). Sustav se gradi na velikoj količini ulaznih tekstova provjerene ispravnosti te se iz tih podataka dobiva frekvencija pojavljivanja n-grama. N-gramski pristup pri analizi rečenice provjerava frekvencije unigrama, bigrama, trigrama, tetragrama i pentagrama u rečenici, pa na temelju njih zasebno ili u okolnom kontekstu radi zaključke o njihovoj ispravnosti. N-grami imaju svojstvo pamćenja redoslijeda pojavljivanja riječi.

Prema [4], pogreške u jeziku dijele se na četiri kategorije:

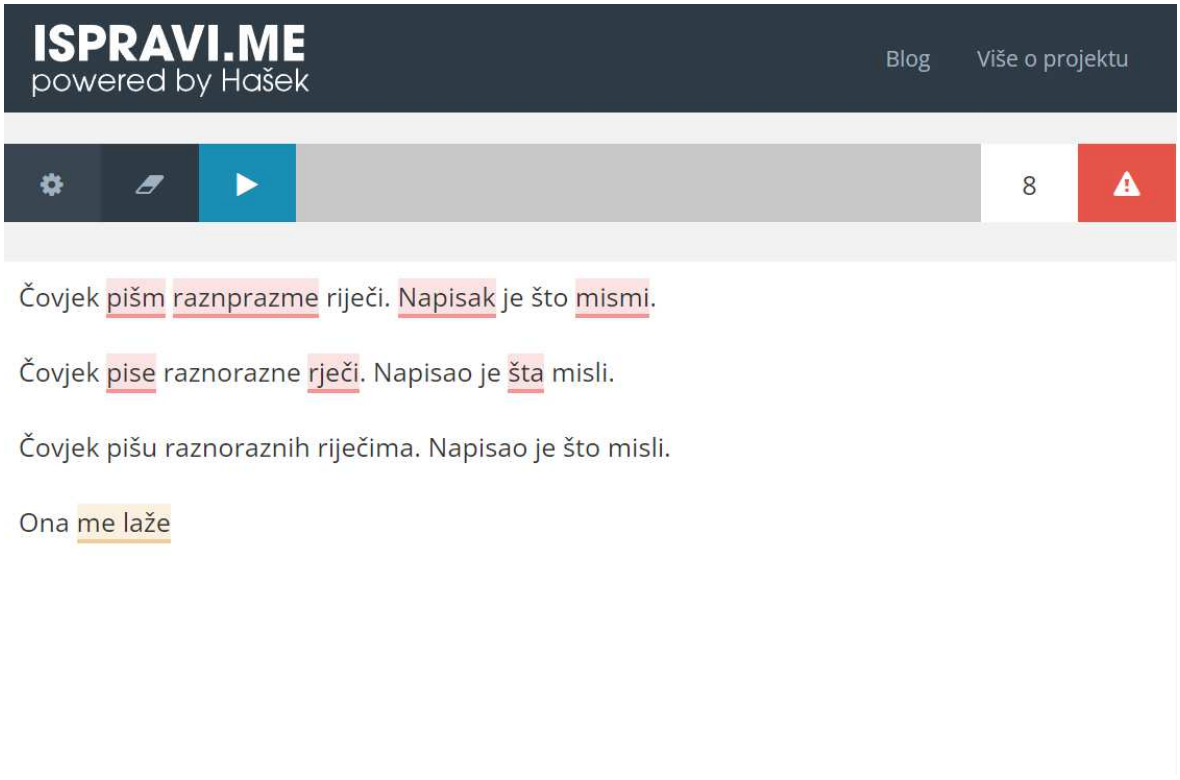
- Pravopisna pogreška, ovom pogreškom nastaje riječ koja nije u jeziku. Za pronalazak pravopisnih pogrešaka potrebno je pogledati nalaze li se riječi u bazi ispravnih riječi, pa za njih nije potrebna kontekstna analiza teksta.
- Gramatička pogreška, pogreška su riječi koje se nalaze u jeziku, ali koriste se na krivome mjestu, protivno pravilima jezika. Za gramatičke pogreške potrebna je kontekstna analiza teksta.
- Stilistička pogreška, ova pogreška predstavlja pogreške nastale korištenjem krivog stila, moguće netipičnim riječima i izrazima.
- Semantička pogreška, to su pogreške koje stvaraju informacije koje su činjenično krive. Te greške teško su uočive, potreban je širok kontekst za njihovo pronalaženje.

Usluga *ispravi.me* primarno ispravlja pravopisne pogreške i zatipke, a uz njih ispravlja i česte gramatičke i stilske pogreške [1]. Stilske pogreške ispravljaju se provjeravanjem postoji li riječ u fondu riječi te, ako ne postoji, nudi sličnopisnice pogrešnoj riječi. Gramatičke i stilske pogreške uočavaju se kontekstnim provjeravanjem preko n-grama i zapisanim pravilima koja govore da su neki sljedovi riječi neispravni.

[5] navodi sljedeće podatke višegodišnje analize rada sustava:

- 18,3 % pogrešaka je nekorištenje dijakritičkih znakova
- 12 % pogreške s *ije* i *je*
- 9,7 % pogreške s *č* i *ć*
- 5,8 % pogreška sa *s* i *sa*

Na slici 1. prikazano je sučelje usluge *ispravi.me* na kojem se demonstriraju trenutne sposobnosti usluge. Vidljivo je da se uočavaju: sve pravopisne pogreške, česte gramatičke greške poput spajanja glagola *lagati* s riječi u akuzativu.



Slika 1. Sučelje usluge *ispravi.me* nakon analize teksta danog na obradu

Promotri li se treći redak na slici 1., uočavaju se gramatičke pogreške: imenica *čovjek* i glagol *pisati* ne slažu se u licu, pridjev *raznoraznih* i imenica *riječima* ne slažu se u padežu. Kao buduće poboljšanje usluge *ispravi.me*, razmatra se razvoj tekstualne analize koja bi

mogla, koristeći fondove riječi i n-gramske sustave, uočavati i ispravljati gramatičke pogreške kako bi korisnici dobili, uz ispravljanje brojnih zatipaka, informaciju pišu li gramatički smislene rečenice.

2. Osnove gramatike

Gramatičke pogreške krše pravila gramatike jezika. Do 19. stoljeća gramatika se uglavnom shvaćala kao skup pravila kojima se propisuje kako se nekim jezikom govori i piše, a od 19. stoljeća gramatika se uglavnom poima kao teorijska i metodološka znanstvena disciplina u kojoj se proučavaju zakonitosti funkcioniranja jezik, napose ustroj i uloga jezičnih jedinica [6]. Grane gramatike su

- Fonologija - dio gramatike koja proučava foneme ili glasove kao razlikovne jedinice [7]
- Morfologija – dio gramatike koji se bavi promjenama oblika riječi u sklonidbi (deklinaciji) i sprezanju (konjugaciji) i promjenama gramatičkih morfema uzrokovanima tim promjenama [8]
- Sintaksa - dio gramatike koji proučava poredak, razmještaj i međusobno slaganje riječi u izraze ili sintagme, rečenične službe i slaganje rečenica u veće cjeline [9]

Gramatičke pogreške uglavnom su pogreške krive uporabe riječi ili njezinog oblika, stavljanje riječi na krivo mjesto u rečenici [3].

Stilistika je filološka disciplina koja proučava jezične stilove [10]. Stilske pogreške ne krše gramatička pravila jezika, ali su u sukobu sa standardnim jezikom i drugim stilovima [11].

Za razvitak usluge koja ispravlja gramatičke pogreške potrebno je znati osnove i gramatička pravila hrvatskoga jezika kako bi se ispravno pristupilo izradi algoritama.

Prema [11] gramatičke kategorije one su kategorije u koje smještamo riječi na temelju njihove tvorbe ili funkcije. Postoje sljedeće kategorije:

1. vrsta riječi – imenice, pridjevi, brojevi, zamjenice, glagoli, prilozi, prijedlozi, čestice, usklici i veznici
2. broj – jednina i množina
3. padež – nominativ, genitiv, dativ, akuzativ, vokativ, lokativ i instrumental
4. rod – muški, ženski i srednji
5. vid – svršeni i nesvršeni
6. stanje – radno i trpno
7. lice – prvo, drugo i treće
8. vrijeme – prošlost, sadašnjost i budućnost
9. način – izjava, zapovijed, mogućnost, želja

10. rečenični član – subjekt, predikat, objekt, priložna oznaka

Gramatičke kategorije roda, broja i padeža vežu se uz imenice, zamjenice, pridjeve i brojeve, a kategorije vida, vremena, načina, roda, stanja i lica uz glagole. Sintagma je spoj dviju punoznačnih riječi. Nastaju preko tri tipa gramatičke veze: sročnosti (kongruencijom), upravljanjem (rekcijom) i pridruživanjem. Sročnost je veza koja nastaje podudaranjem riječi u rodu, broju i padežu, najlakše vidljiva u odnosu pridjeva, brojeva, zamjenica i imenice koju opisuju („jedan ubav srok“). Upravljanje je veza gdje glavna riječ upravlja gramatičkim svojstvima zavisne riječi. Najjasniji je primjer upravljanja način na koji glagoli određuju padež imenice koja ovisi o njima, primjerice predikat (glagol) otvara mjestu objektu (imenici), a ta imenica može biti samo u određenom padežu (akuzativu, dativu, genitivu), ovisno o glagolu („Ja želim parcelu.“). Pridruživanje je veza gdje zavisna riječ nije promjenjiva i pridružuje se glavnoj riječi. Najčešće je riječ o vezi glagola i priloga („razmišljam pristrano“), no ne i isključivo.

2.1. Red riječi u rečenici

Prema [12], hrvatski jezik koristiti se u stilski neobilježenom i stilski obilježenom redu riječi. U stilski neobilježenom redu riječi nijedna riječ u rečenici nije posebno istaknuta te taj stil odiše objektivnošću. U stilski obilježenom redu riječi neke se riječi posebno ističu zbog svoje važnosti za govornika ili autora. Ruši se neutralni neobilježeni poredak riječi i stvara se obrnuti redoslijed koji negdje dodaje, a negdje oduzima naglasak.

U stilsko neobilježenom redu riječi moguće je napraviti neke generalizacije. Promatraju li se osnovni elementi rečenice, predikat, subjekt i objekt, oni se nalaze u redoslijedu subjekt-predikat-objekt.
Priložne oznake izrečene prilogom nalaze se ili ispred predikata ako je predikat u formi jednostavnog glagolskog oblika („Ja rado hodam“) ili između dijelova predikata ako je predikat u formi složenih glagolskih oblika („Ja sam rado hodao“). Ako se priložna oznaka izriče prijedložnim izrazom, pojavljuje se iza predikata („Ja sam hodao šumom“). Pridjevski atributi nalaze se ispred riječi koja im je otvorila mjesto iliti one koju opisuju, a imenski iza. Apozicija se nalazi ispred imenice koju opisuje.

U stilsko obilježenom redu riječi poredak predikata, subjekta, objekta i priložnih oznaka jest slobodan. Redoslijed ovisi o govorniku ili piscu koji želi naglasiti dijelove rečenice, skupine riječi ili same riječi. Najistaknutija mjesta u rečenici jesu prvo i posljednje mjesto.

Postoji i obvezatan red riječi u rečenicama. Riječ je o pravilima koja su obavezna zbog ritmičko-intonacijskih pravila hrvatskoga jezika, o pravilima položaju nenaglasnica – prednaglasnica i zanaglasnica. U daljnjem tekstu navedena su neka od pravila.

Zanaglasnice stoje iza prve naglašene riječi i teže početku rečenice i grupiranju („To **me je** neugodno iznenadilo.“). Ne mogu stajati na početku rečenice, iza niječne čestice i veznika *i*, *a*. U zavisnosloženim rečenicama zanaglasnice se nalaze odmah iza veznika te upitnih i odnosnih zamjenica. Zanaglasnica *je* stoji iza zamjeničkih zanaglasnica. Ako u rečenici stoji više zamjeničkih zanaglasnica jedna do druge, prvo dolazi zanaglasnica u dativu, pa u genitivu, pa u akuzativu. Zanaglasnice mogu biti između pridjevskog atributa i imenice ili dvaju pridjevskih atributa („Drveni **se** stol istrošio.“).

Prednaglasnice se nalaze ispred prve naglašene riječi. Prijedlozi stoje ispred riječi kojoj određuju padež. Prijedlozi stoje između sastavnih dijelova neodređenih zamjenica (ispravno je „ni s kim“, ne „s nikim“).

Postoje još brojna potanko opisana pravila koja se ovdje ne navode zbog manje primjenjivosti tematici rada i kako se ne bi ulazilo u nepotrebne detalje.

2.2. Funkcionalni stilovi hrvatskoga standardnog jezika

Prema [11], hrvatski standardni jezik jest polifunkcionalan. On se na različite načine koristi u znanosti, novinarstvu, administraciji, književnosti i u razgovoru. Po tome postoji pet funkcionalnih stilova hrvatskoga jezika: znanstveni funkcionalni stil, administrativno-poslovni funkcionalni stil, novinarsko-publicistički funkcionalni stil, književnoumjetnički funkcionalni stil i razgovorni funkcionalni stil. Svaki stil ima neka svoja pravila koja ga izdvajaju od drugih stilova ili čak stavljaju u opreku s drugim stilovima.

Znanstveni funkcionalni stil je stil koji je objektivan. U tom stilu apstraktnost, udaljenost od pisca, depersonaliziranje, postiže korištenjem, primjerice, trećeg lica prezenta. Kategorija vremena nije bitna, pa se koristi prezent glagola koji je neutralan, jer taj prezent u tekstu označava prošlost, sadašnjost i budućnost. Rečenice u znanstvenom stilu samostalne su u smislu da imaju i subjekt, i predikat, i objekt, i priložne oznake. Znanstvenom stilu više odgovaraju jednoznačne riječi. Zato je taj stil sklon korištenju internacionalizama – latinizama, grecizama.

Administrativno-poslovni funkcionalni stil jednostavan je, jasan, točan, kratak, stilski neobilježen i sklon učestalom korištenju nekih izraza i sintagmi.

Novinarsko-publicistički funkcionalni stil koristi se u novinarstvu i publicistici. Očekuje se da je on jezično i stilski raznolik, no hrvatsko novinarstvo pokazuje da je u hrvatskome jeziku novinarsko-publicistički stil često blizak s administrativno-poslovnim stilom [11].

Književnoumjetnički funkcionalni stil poseban je stil jer se on ne može karakterizirati putem normi o jeziku. Stil ne koristi jezična pravila kako bi znao kako se nešto treba izgovoriti, nego koristi i izvrće pravila kako bi iznio ono što želi izgovoriti.

Razgovorni funkcionalni stil koristi se u svakodnevnoj komunikaciji. Stil koristi barbarizme, dijalektizme, regionalizme, vulgarizme. Nisu mu strane eliptične rečenice, krnji infinitivi, poštapalice, redoslijed je komponenata slobodan.

3. Opis zadatka i cilja rada

Svrha ovog rada je analizirati i nadograditi leksičke baze usluge *ispravi.me* i razraditi sustav za ispravljanje gramatičkih i drugih kontekstualnih pogrešaka u tekstu kako bi se nadogradila funkcionalnost usluge.

Trenutno, leksički fondovi usluge *ispravi.me* sadrže riječi bez informacija o toj riječi. Fondovi, a i n-gramska baza, ne govore ništa o tome što te riječi u njima jesu, kakva gramatička svojstva imaju.

Kako bi bila moguća gramatička analiza teksta, potrebno je riječima pridodati gramatičke kategorije koje bi se koristile za računanje gramatičke ispravnosti teksta, odnosno krše li riječi u rečenici neko ili neka gramatička pravila. Stoga je potrebno leksičke baze usluge *ispravi.me* nadopuniti s podatcima o gramatičkim kategorijama riječi. Gramatičke kategorije mogu se odrediti na nekoliko načina [13]:

- proučavanjem morfema riječi, na primjer, ako riječ završava na *-nost*, onda je vjerojatno imenica muškog roda u nominativu jednine
- proučavanjem sintagmi, međuodnosa riječi, na primjer, ako odredimo da je nešto prijedlog *u*, onda nakon tog prijedloga slijedi imenica ili njezini atributi u lokativu

To nije jasan algoritam što se lako demonstrira primjerom. Ako bi se riječi *smislen* i *jasen* probale kategorizirati po završetku *-en* dobili bismo da je *jasen* pridjev ili *smislen* imenica. Ovaj primjer pokazuje da su preklapanja morfema različitih kategorija velika, a postoje i primjeri gdje se riječi identično pišu, no različite su vrste riječi, poput pridjeva *brzo* i priloga *brzo*.

Nakon što bi se riječima pridodale gramatičke kategorije, slijedilo bi korištenje jezičnih pravila i gramatičkih kategorija riječi za uočavanje gramatičkih pogrešaka. Dakle, cilj je omogućiti da se u sustav upiše rečenica, da sustav pridijeli gramatičke kategorije riječima i da se preko gramatičkih pravila, a i kontekstualne n-gramske analize, odredi postoje li gramatičke pogreške. Kontekstualna n-gramska analiza korisna je jer je n-gramska baza građena na velikim korpusima riječi i tekstovima korisnika. Preko n-grama je stoga nekad lakše uočiti grešku vjerojatnosnim pristupom nego razmrsiti neku kompleksnu rečenicu.

4. Označavanje gramatičkih kategorija

Označavanje gramatičkih kategorija (engl. *Part of Speech Tagging*) postupak je pridodavanja gramatičkih oznaka riječima u tekstu temeljeno na jezičnim pravilima, statistici i/ili strojnim učenjem [14]. Čistom ulaznom tekstu pridodaju se gramatičke oznake koje točno opisuju odnose i stanja riječi u rečenicama. Označavanje gramatičkih kategorija proces je razjašnjenja višeznačnosti koje tekst sadrži: u hrvatskome je jeziku, kao i u ostalim jezicima, zapisivanjem riječi došlo do istopisnica, odnosno homografa. Postoje riječi koje se izgovaraju identično te se tako i pišu identično, poput glagola *ìzbavi* (imperativ glagola *izbaviti*) i *ìzbavi* (aorist glagola *izbaviti*); neke se riječi izgovaraju različito, zbog naglasaka, no postaju istopisnice jer se u tekstu ne bilježe naglasci, poput pridjeva *sâm* i glagola *sam*. Istopisnice su glavni problem označavanja gramatičkih kategorija riječi. Razlog su tome da se za neku pisanu riječ ne može pridodati određena gramatička kategorija ako postoje druge riječi koje se pišu isto. U tom slučaju mora se odlučiti koju gramatičku kategoriju pridodati.

Jedan način odlučivanja jest korištenje učestalosti pojavljivanja riječi. Pretpostavimo da postoji sustav koji bilježi frekvencije pojavljivanja svih riječi. Potom, ako se u rečenici analizira riječ za koju postoji više mogućnosti, na primjer, već spomenuti *sâm* i *sam*, odabire se riječ koja se najčešće pojavljuje u skupu podataka na kojemu se baza riječi uči. Ako se dvoumi između riječi *opíjati* (infinitiv glagola) i *opijáti* (nominativ množine imenice), a *opíjati* je učestaliji, riječ se označava gramatičkim kategorijama riječi *opíjati*. Ovakav pristup u engleskome jeziku ima točnost od oko 92 %, dok najsuvremeniji algoritmi imaju točnost od oko 97 % [15]. Bitno je napomenuti da se ovdje radi samo o određivanju vrste riječi, bez analiziranja kompliciranijih gramatičkih kategorija poput padeža, kojih nema u engleskome jeziku. Primjerice, 10 milijuna riječnih oznaka u turskome jeziku sadrži četiri puta više riječi (u smislu riječi s jedinstvenim gramatičkim kategorijama) nego u engleskome [15]. Razlog tomu je to što engleski u najvećoj mjeri nije flektivni jezik, odnosno jezik u kojem se riječi mijenjaju dodavanjem različitih gramatičkih elemenata na osnovu (kod imenica deklinacija, kod glagola konjugacija). Riječi u jezicima koji su pretežno flektivni sadrže više informacija unutar sebe, dok engleski te informacije nadoknađuje s drugim riječima ili ih jednostavno izostavi. Ako riječi sadrže manje informacija, postoji manje riječi koje se isto pišu, a drugačije znače.

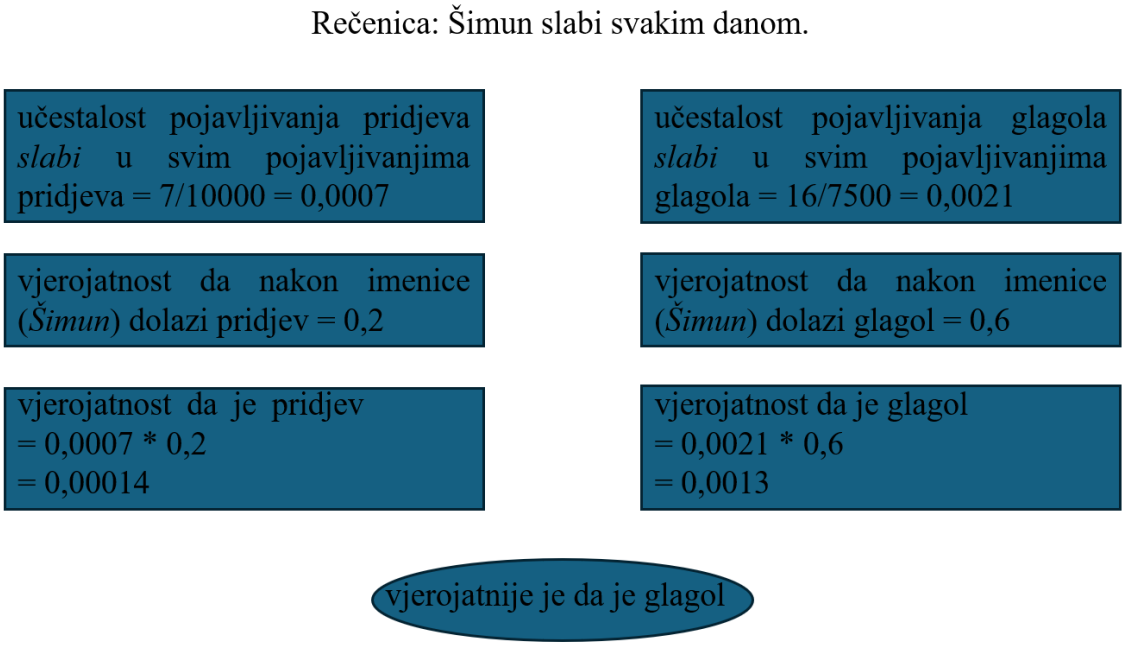
Drugi način određivanja gramatičkih kategorija temelji se na razvijanju algoritama koji u obzir uzimaju kontekst riječi kojoj se određuju svojstva, odnosno gdje se u rečenici nalazi

riječ [15]. U ovome pristupu bitno je razaznati kakvu ulogu riječi imaju u rečenici u jeziku koji se analizira, te način na koji se te uloge međusobno nadopunjavaju. Promotri li se, primjerice, hrvatski jezik, moguće je proučiti međudjelovanje subjekta, glagola i objekta u rečenici, kako rod imenice određuje rod pridjeva te kojim padežima glagol otvara mjesto. Kada se razvije razumijevanje rečeničnih svojstava, može se razvijati algoritam koji, prateći ugrađena pravila, računa vjerojatnosti da neka riječ ima određenu gramatičku kategoriju.

Dva popularna statistička modela koji u obzir uzimaju kontekst riječi kojoj se određuju svojstva su skriveni Markovljevi lanci i uvjetna slučajna polja (engl. *Conditional random fields*, CRF) [15].

Prema [16] Markovljev je lanac „u teoriji vjerojatnosti i statistici niz stanja sustava u kojem stanje u nekom trenutku ovisi samo o stanju u prethodnome trenutku. Uvjetna vjerojatnost da se proces u nekome trenutku u budućnosti nalazi u određenome stanju, ako se u sadašnjosti i prošlosti nalazio u određenome nizu stanja, jednaka uvjetnoj vjerojatnosti istoga budućeg stanja uz uvjet samo sadašnjega stanja.“ Korištenje skrivenih Markovljevih lanaca u označavanja gramatičkih kategorija riječi statistički je model koji se odvija tako da se prvo na skupu podataka za učenje odrede vjerojatnosti prijelaza iz jedne gramatičke kategorije u drugu, primjerice da nakon imenice u nominativu dolazi glagol. Zatim, na temelju skupa podataka za učenje računaju se emisijske vjerojatnosti, vjerojatnosti da će riječ biti generirana iz određene gramatičke kategorije. Primjerice, ako se pridjev *slabi* pojavljuje sedam puta u korpusu, a postoji 10000 pridjeva u korpusu, emisijska vjerojatnost da je pridjev *slabi* jest 7/10000. Kada se nakon učenja analizira neki novi tekst te se dođe do riječi *slabi*, sustav će morati odrediti radi li se o pridjevu ili glagolu. Izračunat će se dvije vjerojatnosti: prva je vjerojatnost umnožak vjerojatnosti da nakon prijašnje riječi dolazi glagol i emisijske vjerojatnosti za glagol *slabi*, druga je vjerojatnost umnožak vjerojatnosti da nakon prijašnje riječi dolazi pridjev i emisijske vjerojatnosti za pridjev *slabi*. Odabire se gramatička kategorija s većom vjerojatnošću (slika 2.). Skriveni Markovljevi lanci gledaju

samo riječ neposredno iza riječi koje se analizira, pa im nedostaje svijest o široj okolini i kontekstu rečenice.



Slika 2. Prikaz procesa računanja vjerojatnosti je li *slabi* glagol ili pridjev

Uvjetna slučajna polja statistički je model kojemu možemo pridružiti proizvoljne uvjete za određivanje gramatičkih kategorija riječi. Na primjer, može se napraviti sustav za označavanje gramatičkih oznaka koji će određivati riječ na temelju: kategorija riječi prije, završetka riječi koju analiziramo, kategorija riječi nakon, broja riječi u rečenici. Te su značajke parametri uvjetnih slučajnih polja. Prednost uvjetnih slučajnih polja naspram skrivenih Markovljevih lanaca je mogućnost korištenja više značajki, dok skriveni Markovljev lanac koristi samo riječ prije riječi koju analiziramo. Također, uvjetna slučajna polja mogu računati, odnosno pretpostavljati, gramatičke kategorije nepoznatih riječi, na primjer preko morfoloških karakteristika [15].

5. Pristup i metodologija dodavanja gramatičkih kategorija riječima

Na dva se načina može pristupiti dodavanju gramatičkih kategorija riječima iz baze usluge *ispravi.me* kako bi se omogućio razvoj algoritma za označavanje gramatičkih kategorija riječi u tekstu:

1. Započeti novu analizu riječi i njihovih svojstava. To znači poluautomatizirano pridodavanje gramatičkih svojstava riječima. Svakako, postoje uzorci koji se mogu generalizirati, no svaki jezik ima svoje iznimke i nepravilnosti koji su teški za ukalupiti.
2. Nastaviti se na tuđi rad pridodavanja gramatičkih svojstava riječima. Na primjer, Hrvatski jezični portal sadrži osnovne podatke riječi u sustavu, sve oblike glagola, imenica, pridjeva, zamjenica.

Ideja ovog rada je okoristiti se javno dostupnim servisima, odnosno koristiti drugi pristup.

hrLex leksikon je hrvatskoga jezika javno dostupan za korištenje. Prva verzija leksikona objavljena je 5. ožujka 2016. godine, a treća 31. ožujka 2019. godine. U radu se koristi treća verzija. Autori leksikona su Nikola Ljubešić, Filip Klubička, Željko Agić i Ivo-Pavao Jazbec [17][18]. Leksikon sadrži informacije o 164 206 hrvatskih lema riječi. Lema riječi je kanonski oblik riječi, primjerice za glagol je to infinitiv, pa glagoli *vidim*, *vidjesmo*, *vidi* imaju istu lemu, *vidjeti* [19]. Nisu u leksikonu sadržane samo leme riječi, već i drugi oblici riječi koje imaju tu lemu. Leksikon ima 6 427 709 unosa. Svaki od unosa je jedinstvena riječ po slovima i njezinim gramatičkim kategorijama.

hrLex označava riječi gramatičkim oznakama po standardu *MULTEXT-East Morphosyntactic Specifications*. Navedeni standard koristi se kako bi se osiguralo korištenje istih morfosintaktičkih specifikacija za sljedeće jezike: albanski, bugarski, čečenski, češki, estonski, engleski, gruzijski, hrvatski, mađarski, makedonski, perzijski, poljski, rumunjski, ruski, srpski, slovački, slovenski i ukrajinski [20].

Svaki redak *hrLexa* sastoji se od osam izraza razmaknutih tabulatorom:

- riječ
- lema te riječi
- morfosintaktička specifikacija
- morfosintaktička svojstva

- UPOS (engl. *Universal Part Of Speech*), oznaka gramatičke kategorije, primjerice ADJ označava pridjev (engl. *Adjective*)
- morfološka svojstva
- broj pojavljivanja riječi
- frekvencija pojavljivanja riječi

Primjera radi, prikazana je osmorka riječi *ječam* u nominativu.

1. ječam
2. ječam
3. Ncmsn
4. Type=common|Gender=masculine|Number=singular|Case=nominative
5. NOUN
6. Case=Nom|Gender=Masc|Number=Sing
7. 365
8. 0.000261

Značenje morfosintaktičke specifikacije razumije se preko legende prikazane u tablicama (Tablica 1., Tablica 2.) [19].

Tablica 1. Kategorije i oznake

Kategorija	**Oznaka**
imenica (engl. *Noun*)	N
glagol (engl. *Verb*)	V
pridjev (engl. *Adjective*)	A
zamjenica (engl. *Pronoun*)	P
prilog (engl. *Adverb*)	R
prijedlog (engl. *Adposition*)	S
veznik (engl. *Conjuction*)	C
broj (engl. *Numeral*)	M
čestica (engl. *Particle*)	Q
usklik (engl. *Interjection*)	I

Tablica 2. Popis atributa svih kategorija riječi

Kategorija	Atribut	Vrijednost atributa
imenica (N)	tip	opća (c), vlastita (p)
	rod	muški (m), ženski (f), srednji (n)
	broj	jednina (s), množina (p)
	padež	nominativ (n), genitiv (g), dativ (d), akuzativ (a), vokativ (v), lokativ (l), instrumental (i)
glagol (V)	tip	glavni (m), pomoćni (a)
	oblik	indikativ (i), imperativ (m), kondicional (c), infinitiv (n), particip (p)
	vrijeme	prezent (p), imperfekt (i), futur (f), perfekt (s), pluskvamperfekt (l), aorist (a)
	lice	prvo (1), drugo (2), treće (3)
	broj	jednina (s), množina (p)
	rod	muški (m), ženski (f), srednji (n)
	stanje	radno (a), trpno (p)
pridjev (A)	tip	opisni/gradivni (f), posvojni (s)
	stupanj	pozitiv (p), komparativ (c), superlativ (s)
	rod	muški (m), ženski (f), srednji (n)
	broj	jednina (s), množina (p)
	padež	nominativ (n), genitiv (g), dativ (d), akuzativ (a), vokativ (v), lokativ (l), instrumental (i)
	određenost	neodređen (n), određen (y)
zamjenica (P)	tip	osobna (p), pokazna (d), posvojna (s), upitna (q), neodređena (i), povratna (x)
	lice	prvo (1), drugo (2), treće (3)
	rod	muški (m), ženski (f), srednji (n)
	broj	jednina (s), množina (p)

	padež	nominativ (n), genitiv (g), dativ (d), akuzativ (a), vokativ (v), lokativ (l), instrumental (i)
	referenca	osobna (p), posvojna (s)
	sintaksni tip	imenski (n), pridjevski (a)
prilog (R)	tip	opći (g)
	stupanj	pozitiv (p), komparativ (c), superlativ (s)
prijedlog (S)	tip	opći (p)
	složenost	jednostavni (s), složeni (c)
	padež	nominativ (n), genitiv (g), dativ (d), akuzativ (a), vokativ (v), lokativ (l), instrumental (i)
veznik (C)	tip	nezavisni (c), zavisni (s)
	složenost	jednostavni (s), složeni (c)
broj (M)	tip	glavni (c), redni (o)
	rod	muški (m), ženski (f), srednji (n)
	padež	nominativ (n), genitiv (g), dativ (d), akuzativ (a), vokativ (v), lokativ (l), instrumental (i)
	oblik	znamenka (d), rimska znamenka (r), slovo (l)
čestica (Q)	tip	negativna (z), upitna (q), modalna (o), potvrdna (r)
usklik (I)	složenost	jednostavni (s), složeni (c)

Analizira li se *hrLex* baza po kategorijama vrste riječi, vidi se da najveći broj riječi u bazi čine pridjevi, njih čak 64 %, a za njima slijede vlastite imenice, kojih ima malo manje od 24 %, imenice sa 7,7 %, te glagoli s 3,5 %. Točna raspodjela po svim UPOS oznakama vidljiva je u tablici 3.

Tablica 3. Raspodjela vrsta riječi u hrLexu

UPOS oznaka	Broj pojavljivanja	Postotak u bazi
pridjevi (ADJ)	4 123 591	64,153
vlastite imenice (PROPN)	1 536 132	23,898
opće imenice (NOUN)	496 832	7,729
glagoli (VERB)	223 664	3,479
prilozi (ADV)	42 876	0,667
odrednice (DET)	3 156	0,049
brojevi (NUM)	775	0,012
zamjenice (PRON)	238	0,004
prijedlozi (ADP)	109	0,002
uzvici (INTJ)	106	0,002
pomoćni glagoli (AUX)	66	0,001
čestice (PART)	41	0,001
podredni veznici (SCONJ)	27	0,000
koordinacijski veznici (CCONJ)	17	0,000
interpunkcijski znakovi (PUNCT)	11	0,000

U pomoćne glagole spadaju glagoli biti i htjeti. Neke manje kategorije nisu naširoko poznate i korištene u hrvatskom razgovornom jeziku, no, kako je riječ o međunarodnom standardu POS oznaka, te oznake koriste se i u *hrLexu*.

Brojčanu dominaciju pridjeva u bazi riječi ovog tipa objašnjava činjenica da se pridjevi mogu nalaziti u tri roda, dok imenice samo u jednom. Postoji sedam padeža i dvije kategorije broja (jednina i množina). Svaka kombinacija te dvije kategorije daje jedan redak, pa tako imenice generiraju 14 redaka, dok pridjevi imaju sedam padeža, dvije kategorije broja, tri roda, tri stupnja (pozitiv, komparativ, superlativ) i opreku određenost/neodređenost. Na primjer, imenica *smislu* imat će u *hrLexu* dva retka:

- prvi za riječ u jednini muškog roda u dativu
- drugi za riječ jednine muškog roda u lokativu

zato što riječi u dativu i lokativu imaju isti oblik. Ako se pogleda pridjev *smislenomu*, on ima četiri retka:

- muški rod jednine u dativu
- muški rod jednine u lokativu
- srednji rod jednine u dativu
- srednji rod jednine u lokativu.

Iako nemaju svi pridjevi sva ova svojstva, mnogi ih imaju, a time svaka lema pridjeva može generirati preko 200 redaka.

U bazi postoji 8.3 puta više pridjeva nego općih imenica, a lema pridjeva generira preko 12 puta više redaka nego lema opće imenice.

Baza *hrLex* sadrži mnogo prezimena koja se kategoriziraju kao vlastite imenice. I tako, budući da sustav kategorizira prezimena i po muškom i po ženskom rodu, generiraju dvostruko više redaka nego opće imenice.

6. Pokrivanje jezične baze

Kako bi se za rječničku bazu usluge *ispravi.me* razvila nekakva vrsta analize teksta koristeći gramatičke kategorije i jezične zakonitosti, potrebno je pretvoriti bazu usluge *ispravi.me* u bazu nalik *hrLexu* – riječima dodati njihove gramatičke kategorije i omogućiti iz homonimija razaznavanje riječi koje su različite, bilo u vrsti riječi, bilo u nekoj drugoj gramatičkoj kategoriji. Riječ iz baze usluge *ispravi.me* generirala bi jednu ili više riječi u novoj bazi jer se riječi u bazi usluge *ispravi.me* nalazi potencijalno više istopisnica. Za riječ *sam* generirat će se redci:

- glagol biti
- pridjev muški rod jednine nominativ
- pridjev muški rod jednine akuzativ
- pridjev muški rod jednine vokativ.

Time bi se povećao broj riječi u bazi te riječi ne bi bile, kao dosad, u sustavu bez ikakvih podataka o prirodi riječi.

Osnovna je ideja riječima iz baze usluge *ispravi.me* naći parove u bazi *hrLex* te preslikati svojstva tih riječi, gramatičke kategorije, u bazu usluge *ispravi.me*.

Napisan je kod u programskom jeziku Rust koji za svaku riječ u općejezičnom fondu traži redak u *hrLexu*, a ako ih ima više, odabire redak s najvećom učestalosti. Svrha ovog programa jest da se sazna koliko je riječi iz baze usluge *ispravi.me* pokriveno, odnosno postoji li barem jedan redak u *hrLexu* kojem je izraz isti onom u općejezičnom fondu.

Pseudokod 1. Izračun početne pokrivenosti baze usluge *ispravi.me*

```
1   hrLex ← učitaj hrLex bazu
2   types ← učitaj bazu općih riječi usluge ispravi.me
3   izlaz ← nova datoteka u koju se zapisuju rezultati
4   za redakT u types
5       lista ← redci iz hrLexa koji počinju s redakT
6       ako lista != []
7           najčešćaRiječ ←redak iz liste koji ima najveću frekvenciju pojavljivanja
8           zapiši(izlaz, najčešćaRiječ)
9       inače
10          zapiši(izlaz, redakT)
```

11 ***brojPokrivenihRiječi*** ← *broj redaka s više od jedne riječi u* ***izlaz***

12 ***ukupanBrojRiječi*** ← *broj redaka u* ***izlaz***

13 ***pokrivenost*** ← ***brojPokrivenihRiječi/ukupanBrojRiječi***

U Pseudokodu 1. raspisan je algoritam pokrivanja.

Dobiveni podaci pokazuju da je pokriveno 50,3 % riječi u općejezičnom fondu, njih 558 677 od 1 110 952.

Postavlja se pitanje zašto su neke riječi u općejezičnom fondu, a nisu u *hrLexu* i obratno. Svakako treba pogledati kada su nastale baze. Baza *hrLex* nastala je 2019. godine, a općejezični fond raste i razvija se kroz godine te raste na temelju tekstova prispjelih na pravopisnu obradu. Posljedica toga je to što je općejezični fond malo ažurniji od *hrLexa* kada se gledaju novonastale riječi i riječi koje su popularne zadnjih godina.

Za dobru analizu rečenica i tekstova nužno je postojanje baze s mnogo riječi, a pokrivenost od 50,3% nije dovoljno velika da bi se moglo koristiti bazu bez da često u sustav dolaze nepoznate riječi. Dakle, nužno je povećati pokrivenost.

Puno riječi iz baze usluge *ispravi.me* nema ekvivalent u bazi *hrLex*. To ne znači da nema daljnje koristi od *hrLexa* jer:

- *hrLex* ima raspisanu kategorizaciju pojmova,
- *hrLex* sadrži mnogo riječi koje su slične nekima u bazi usluge *ispravi.me.*

Kako bi se povećala pokrivenost općejezičnog fonda potrebno je iskoristiti ove dvije točke: za nepokrivene riječi u općejezičnom fondu pronaći ne identične riječi u *hrLexu*, nego riječi koji imaju ekvivalentne gramatičke kategorije.

Na površinu dolaze dvije metode koje ne zahtijevaju razvoj kompleksnih algoritama, nego poznavanje prirode hrvatskog jezika i uočavanje uzoraka koji se pojavljuju u riječima. Kada se uoče uzorci, mogu se preslikati određeni redci iz *hrLexa* u općejezičnom fondu. Prva metoda temelji se na prefiksima, a druga na sufiksima ili završetcima riječi.

6.1. Prefiksacija

Razmotrimo sljedeći slučaj. Nakon prvotne analize pokrivenosti i dobivenih 50,3 %, promatra se koje se riječi nalaze u preostalih 49,7 % i može li se o njima štogod zaključiti.

U općejezičnom fondu pronađena je riječ *subdistribucija*. U *hrLexu* ne postoji *predefiniranost*, no postoji *definiranost*. Razmisli li se o tim dvjema riječima, uviđa se da te riječi imaju identične gramatičke kategorije. Ako u *hrLexu* postoje redci s vrijednostima vidljivima u tablici, dopušteno je preslikati te retke u općejezični fond i stvoriti za *predefiniranost* onoliko redaka koliko *definiranost* ima u *hrLex*.

***hrLex*:**

definiranost definiranost Ncfsa ...

definiranost definiranost Ncfsn ...

općejezični fond:

predefiniranost predefiniranost Ncfsa ...

predefiniranost predefiniranost Ncfsn ...

Analizirajući nepokrivene riječi u općejezičnom fondu i općenito prefikse hrvatskog jezika, popisani su sljedeći prefiksi koji se često nalaze u riječima općejezičnoga fonda (pravopis.hr). Ti se prefiksi koriste u pospješivanju pokrivenosti općejezičnoga fonda:

a-, aero-, agro-, alo-, an-, anti-, auto-, be-, bes-, bez-, beza-, beš-, bi-, bio-, bruto-, di-, dis-, disko-, do-, drugo-, eko-, ekstra-, etno-, euro-, foto-, gastro-, hidro-, hiper-, info-, ino-, inter-, is-, iz-, iza-, između-, izo-, izvan-, jedno-, kino-, ko-, kontra-, krim-, krimi-, kvazi-, latino-, mega-, među-, mikro-, mini-, mono-, multi-, na-, nad-, naj-, nano-, nat-, nazovi-, ne-, nekro-, neo-, neto-, neuro-, nisko-, nuz-, o-, ob-, od-, op-, ot-, paleo-, para-, po-, pod-, poli-, polu-, poslije-, post-, pot-, pra-, pre-, pred-, preko-, pri-, prije-, pro-, promo-, protu-, prvo-, pseudo-, psiho-, radio-, ras-, raz-, raza-, raš-, re-, samo-, ski-, socio-, stereo-, su-, sub-, super-, sve-, taksi-, tara-, trans-, u-, ultra-, uz-, van-, vele-, vice-, video-, visoko-, z-, za-, žiro-

Pseudokod 2. Izračun pokrivenosti baze usluge *ispravi.me* koristeći prefikse

1 ***hrLex*** ← *učitaj hrLex bazu*

2 ***types*** ← *učitaj bazu općih riječi usluge ispravi.me*

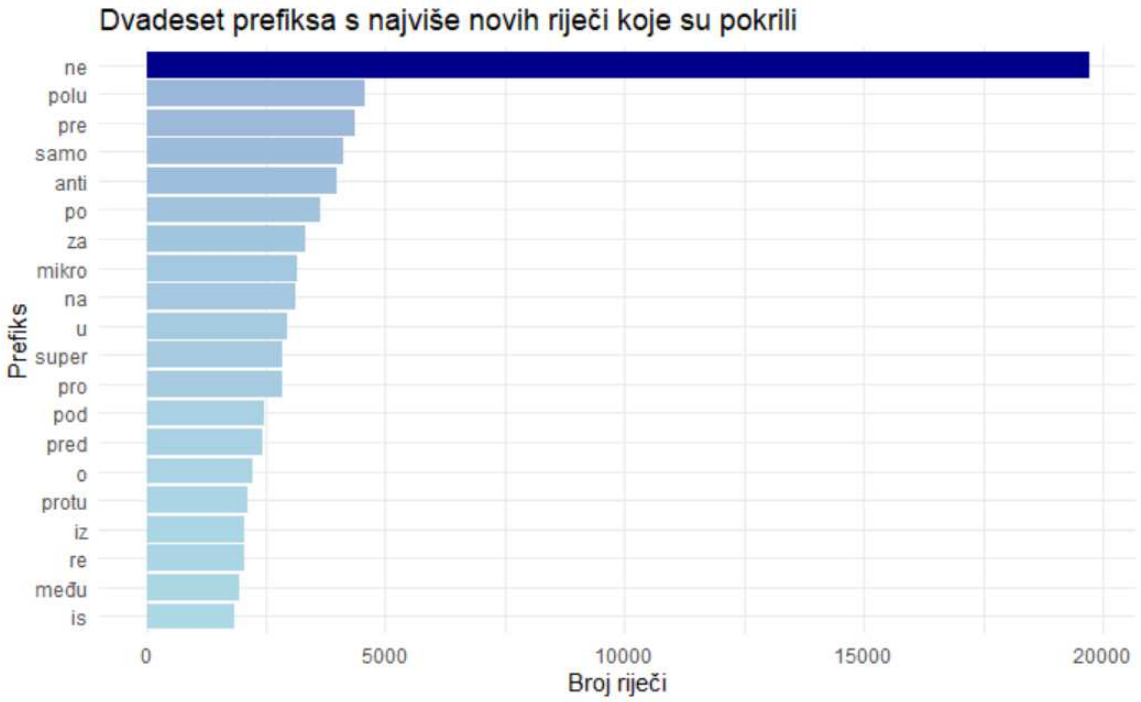
3 ***izlaz*** ← *nova datoteka u koju se zapisuju rezultati*

4 **prefiksi** ← *lista prefiksa koji se koriste u analizi*

```
5   za redakT u types
6       lista ← redci iz hrLexa koji počinju s redakT
7       ako lista != []
8           najčešćaRiječ ← redak iz liste koji ima najveću frekvenciju pojavljivanja
9           zapiši(izlaz, najčešćaRiječ)
10      inače
11          Za prefiks u prefiksi
12              Ako riječKrećeS(redakT, prefiks)
13                  lista ← redci iz hrLexa koji počinju s redakT bez prefiksa
14                  ako lista != []
15                      najčešćaRiječ ← redak iz liste koji ima najveću frekvenciju
                        pojavljivanja
16                      zapiši(izlaz, najčešćaRiječ)
17                      Zaustavi petlju
18          Ako nije nađeno pomoću prefiksa
19              zapiši(izlaz, redakT)
20  brojPokrivenihRiječi ← broj redaka s više od jedne riječi u izlaz
21  ukupanBrojRiječi ← broj redaka u izlaz
22  pokrivenost ← brojPokrivenihRiječi/ukupanBrojRiječi
```

U Pseudokodu 2. raspisan je algoritam pokrivanja koristeći prefikse.

Općejezični fond sadrži mnogo takvih riječi s nekim prefiksom. Kada se primijeni pokrivanje općejezičnoga fonda s traženjem istih redaka i dodatno pokrivanje s prefiksima, pokrije se 686 761 riječ iliti 61,8 % baze, što je 128 084 više pokrivenih riječi, oko 10,5 %, nego kada se ne koriste prefiksi.



Slika 3. Statistika dodanih novih riječi po prefiksu

6.2. Sufiksacija

Nakon pokrivanja početaka riječi prelazimo na drugi kraj. Da bi se preko završetaka riječi mogla provesti neka generalizacija kojom bi se označila gramatička svojstva riječi treba biti oprezniji nego kod prefiksa. Uzme li se, primjerice, završetak *-anje*, kao u riječi *gledanje*, svaka riječ koja završava tim slijedom imat će ista svojstva: imenica srednjeg roda u nominativu, akuzativu ili vokativu jednine. Analogno i za *-anju*, te su riječi imenice srednjeg roda u dativu ili lokativu jednine. Ako se želi napraviti generalizacija riječi koje završavaju na *-ist*, poput *kist*, ispravno bi se krenulo označavanjem riječi koje završavaju na *-ist* s kategorijama muškog roda jednine u nominativu. No, ako bi se htjelo zahvatiti druge padeže i stoga generaliziralo s nastavkom *-ista*, dobilo bi se traženo svojstvo, recimo, genitiva jednine muškog roda, ali bi se i zahvatile kategorije imenica ženskih roda koje završavaju na *-ista*, poput *bista*, *pista*, *lista* te pridjeva poput *čist*. Onda bismo u sustavu za riječ *kista* imali i redak za muški rod genitiva jednine i redak za ženski rod nominativa jednine. Potrebno je, dakle, uočiti koji završetci se ne prelijevaju u druge gramatičke kategorije, nego daju u većini slučajeva jednoznačan rezultat. Primjer krivo odabranog je završetak *-en*, isto bi se tretirale riječi *pismen* i *alergen*, a jedna je pridjev, druga imenica.

Analizom još nepokrivenih riječi u bazi usluge *ispravi.me* pronađeni su sljedeći završetci koji izbjegavaju prije spomenut rizik:

-anja, -anje, -anjem, -anjima, -anju, -avati, -čar, -čara, -čari, -čarima, -čka, -čke, -čki, -čkih, -čkim, -čko, -čkog, -čkoj, -čkom, -čku, -elj, -elja, -elji, -eljica, -eljima, -elju, -enja, -enje, -enjem, -enjima, -enju, -evati, -graf, -grafija, -gram, -ica, -ij, -ija, -ik, -ilan, -ilne, -ilni, -ilnih, -ilnog, -ilnom, -iran, -irana, -irani, -irati, -ist, -ista, -sti, -istima, -istom, -istu, -ivati, -izam, -izama, -izma, -izmom, -izmu, -ić, -išta, -ište, -ištu, -logija, -logije, -logiji, -logijom, -logiju, -metar, -njeti, -nost, -nosti, -nostima, -nošću, -oid, -oida, -oidi, -oidima, -ovati, -ska, -ske, -ski, -skih, -skim, -sko, -skog, -skoj, -skom, -sku, -stva, -stvo, -stvom, -tika, -tor, -tora, -tore, -tori, -torima, -tost, -tostima, -tošću, -uvati, -uvši, -uća, -uće, -ućih, -ućim, -vost, -vosti, -vostima, -vošću, -ška, -ške, -ški, -ških, -škim, -sko, -škog, -škoj, -škom, -šku, -štvima, -štvo, -štvom, -štvu

Pseudokod 3. Izračun pokrivenosti baze usluge *ispravi.me* koristeći prefikse i završetke

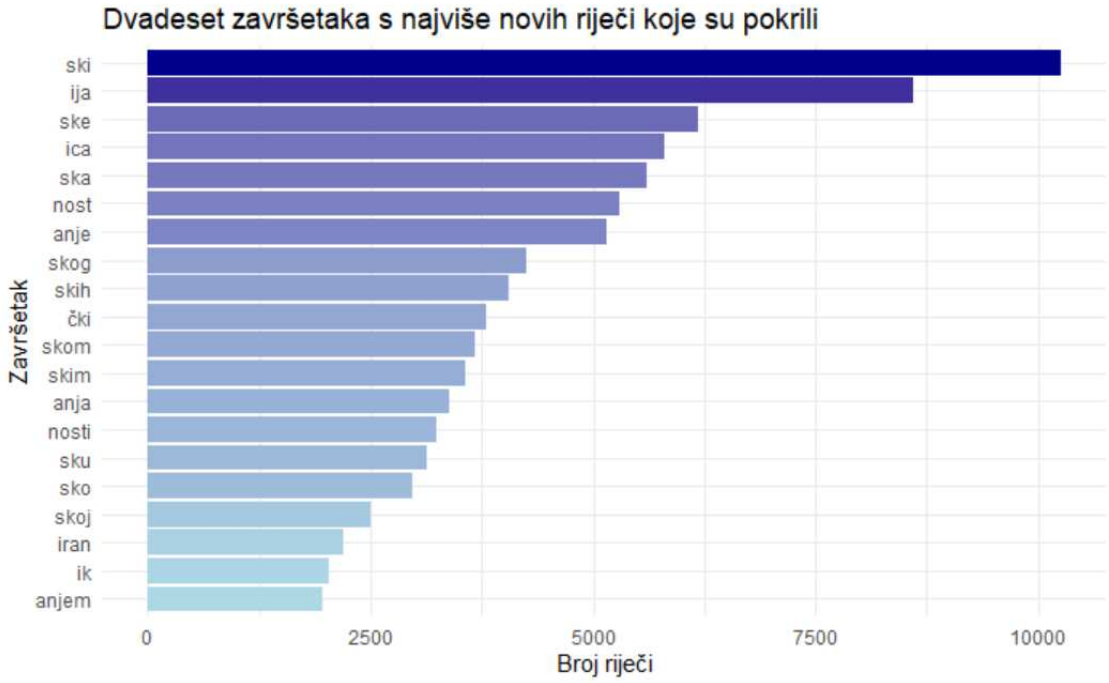
```
1   hrLex ← učitaj hrLex bazu
2   types ← učitaj bazu općih riječi usluge ispravi.me
3   izlaz ← nova datoteka u koju se zapisuju rezultati
4   neNađeni ← riječi koje neće biti pokrivene preko prefiksa
5   prefiksi ← lista prefiksa koji se koriste u analizi
6   sufiksi ← lista završetaka koji se koriste u analizi
7   generalizacije ← datoteka u kojoj se spremaju redci riječi koje završavaju na
    završetak
8   za redakT u types
9       lista ← redci iz hrLexa koji počinju s redakT
10      ako lista != []
11          najčešćaRiječ ← redak iz liste koji ima najveću frekvenciju pojavljivanja
12          zapiši(izlaz, najčešćaRiječ)
13          za sufiks u sufiksi
14              ako riječZavršavaS(redakT, sufiks)
15                  dodajUGeneralizacije(lista, sufiks)
16      inače
17          za prefiks u prefiksi
18              ako riječKrećeS(redakT, prefiks)
```

```
19                      lista ← redci iz hrLexa koji počinju s redakT bez prefiksa
20                      ako lista != []
21                          najčešćaRiječ ← redak iz liste koji ima najveću frekvenciju
                            pojavljivanja
22                          zapiši(izlaz, najčešćaRiječ)
23                          zaustavi petlju
24              ako nije nađeno s pomoću prefiksa
25                  dodaj(neNađeni, redakT)
26  za redakN u neNađeni
27          ako riječZavršavaNaSufiks(redakN)
28              pokrijRiječPrekoRedakaUGeneraliziciji(redakN)
29  brojPokrivenihRiječi ← broj redaka s više od jedne riječi u izlaz
30  ukupanBrojRiječi ← broj redaka u izlaz
31  pokrivenost ← brojPokrivenihRiječi/ukupanBrojRiječi
```

U Pseudokodu 2. raspisan je algoritam pokrivanja koristeći prefikse i završetke.

Kada se u pokrivanju riječi iskoristi i analiza završetaka riječi, pokriveno je sveukupno 829 637 riječi, odnosno 74,7 %. To je 142 876 riječi više nego u pokrivanju preko istih riječi i pokrivanju preko prefiksa.

Pokrivanje preko prefiksa i završetaka nije gotov posao. Postoji još mnogo prefiksa i završetaka koji bi se mogli koristiti u pokrivanju, no oni češći uglavnom su iskorišteni, a preostali pokrivaju skup od sto do dvjesto riječi, ako i toliko. Uz iscrpnu analizu pokrilo bi se, vjerojatno, još nekoliko desetaka tisuća riječi.



Slika 4. Statistika dodanih novih riječi po završetku

Otvara se problematika što učiniti s riječima koje nisu pokrivene niti prefiksima niti završetcima. Nasreću, kako *hrLex* dobro pokriva riječi koje se često koriste, riječi koje nisu pokrivene jesu one koje se zaista rjeđe upotrebljavaju te će se njihov nedostatak manje osjećati. Slijedi jedan nasumično odabrani uzorak nepokrivenih riječi.

osteoplastičnih, arktoalpskomu, ciklobarbiton, adenindinukleotida, gušavice, trinaestooptužena, slikocrteži, frotirastu, transfazor, holokristalast, epiduralnog, merdža, ranopovijesni, vazokonstriktornim, ljubaznice, strignu, mamoplastike, nodalizacijom, odzivnih, klostridijalne, koronamjesec, bolničnog, miorelaksansima, narkosindikat, tehnoentuzijazma, ureterovezikalnom, nederivatna, siktavima, klamidofiloze, neukvasana, osmanistike, svinjogojčev, vračarev, skadencar, mljez, sjevernojužne, upuštača, ponašajnim, okajnicama, homoplazije, snebi, baksuznih, bespolce, osuđivanije, verižac, nonicama, prejudicijelan, akademiziranih, kaučukirane, hlepnju, animatronici, negluhost, hrastovcima, suvlaki, rusoglav, eksaskalarnog, tolvajem, usukanicima, tridesetsedmogodišnjom, zakvačuju, kordicepsu, doksa, srčak, dozive, prepakirač, sitotiskari, orbitnim, raspržniku, utiliziranim, slakomi, kačkane, dropčina...

7. Programska analiza rečenice

Cilj programske analize rečenica nije dosegnut obimom ovog rada, no ostaje prostora za postavljanje temelja takve analize. Kako bi se demonstrirao potencijal ove analize, napisan je program koji vrši jednostavnu analizu nad tekstom koji zaprima. Analiza se sastoji od pridjeljivanja riječima najvjerojatnije vrijednosti gramatičkih kategorija na temelju frekvenciji pojavljivanja koja su uzeta iz *hrLexa*. Nadalje, analizira se nalaze li se neki pridjev i neka imenica jedan do drugoga. Ako je to slučaj, program analizira podudaraju li se te riječi u rodu, broju i padežu, što bi po sročnosti trebalo vrijediti, uz iznimke. Kada se dogodi da program otkrije diskrepanciju u navedenim gramatičkim kategorijama, on javlja da je došlo do moguće greške negdje u tim dvjema riječima. Moguće je da do greške nije došlo, možda su se jedan do drugoga našli pridjev od imenice i neka druga imenica, no najčešće je slučaj da je pridjev napisan kraj riječi koju opisuje. Primjerice, u rečenici „Samo na radosna čovjeka možeš računati“ pridjev *radosna* opisuje imenicu *čovjeka*, a u rečenici „Samo na isprobano čovjek može računati“ pridjev *isprobano* i imenica *čovjek* su jedan do drugoga, no *isprobano* ne opisuje imenicu *čovjek*.

Za potrebu ove analize napravljen je program (Pseudokod 4.) koji na početku učitava novonastalu bazu riječi usluge *ispravi.me* koja sadrži gramatičke kategorije riječi. Zatim se pokreće server koji prima HTTP POST zahtjeve koji sadrže tekst. Taj se tekst šalje u Perl skriptu koja dijeli tekst na rečenice. Za svaku rečenicu koju skripta vrati program je analizira koristeći bazu riječi i riječima pridodaje gramatičke kategorije. Nakon što to dovrši, provjerava nalaze li se pridjevi i imenice jedno kraj drugog i vrši spomenutu analizu. Uz ove informacije ispisuju se za svaku riječ najvjerojatnije gramatičke kategorije koristeći frekvencije pojavljivanja, a i ostale moguće gramatičke kategorije za koje je sustav odlučio da su manje vjerojatne, ali moguće gramatičke kategorije te riječi.

Pseudokod 4. Jednostavna analiza rečenica

```
1   bazaRiječi ← učitaj bazu riječi usluge ispravi.me
2   za tekst
3       rečenice ← perlSkriptaZaPodijeluTekstaNaRečenice(tekst)
4       za rečenicu u rečenice
5           kategorije ← lista u kojoj se spremaju gramatičke kategorije svake riječi u
            rečenici
```

```
6       za riječ u rečenica
7           lista ← redciKojiPočinjuS(bazaRiječi, riječ)
8           najčešći ← redak iz liste koji ima najveću frekvenciju pojavljivanja
9           dodajUKategorije(najčešći)
10          ispiši(lista)
11      za kategorijaRiječi u kategorije
12          ako je riječ pridjev
13              sljedećaRiječ ← kategorije riječi u rečenici nakon trenutno
                analizirane
14              ako je sljedećaRiječ imenica
15                  ispravno ← usporediGramatičkeKategorije(kategorijaRiječi,
                    sljedećaRiječ)
16                  ako != ispravno
17                      ispiši(„Moguće nepodudaranje“, kategorijaRiječi,
                        sljedećaRiječ)
```

```
Učitana baza, šalji!
Primio: vidim pjesnikovu čežnja
vidim pjesnikovu čežnja
  Najvjerojatnija vrijednost: vidjeti   Vmr1s   Type=main|VForm=present|Person=first|Number=singular    VERB    Mood=Ind|Number
  Druge moguće vrijednosti:

  Najvjerojatnija vrijednost: pjesnikov Aspfsay Type=possessive|Degree=positive|Gender=feminine|Number=singular|Case=accusative
s|Gender=Fem|Number=Sing|Poss=Yes       71      0.000051
  Druge moguće vrijednosti:
    pjesnikov   Aspmsdn Type=possessive|Degree=positive|Gender=masculine|Number=singular|Case=dative|Definiteness=no    ADJ
es      0       0.000000
    pjesnikov   Aspmsln Type=possessive|Degree=positive|Gender=masculine|Number=singular|Case=locative|Definiteness=no  ADJ
es      18      0.000013
    pjesnikov   Aspnsdn Type=possessive|Degree=positive|Gender=neuter|Number=singular|Case=dative|Definiteness=no       ADJ
es      0       0.000000
    pjesnikov   Aspnsln Type=possessive|Degree=positive|Gender=neuter|Number=singular|Case=locative|Definiteness=no     ADJ
es      0       0.000000

  Najvjerojatnija vrijednost: čežnja    Ncfsn   Type=common|Gender=feminine|Number=singular|Case=nominative     NOUN    Case=No
  Druge moguće vrijednosti:
    čežnja      Ncfpg   Type=common|Gender=feminine|Number=plural|Case=genitive NOUN    Case=Gen|Gender=Fem|Number=Plur 63


Nađen niz pridjev-imenica: pjesnikovu čežnja
  Nepodudaranje u padežu: padež pridjeva: Case=accusative, padež imenice: Case=nominative
```

Slika 5. Ispis analize rečenice

Slika 5. prikazuje ispis programa pri analizi rečenice „Vidim pjesnikovu čežnja“.

Ovaj primjer demonstrirao je označavanje gramatičkih oznaka isključivo na temelju frekvencije pojavljivanja. No, ovaj sustav ima manjkavosti. Ako se na obradu pošalje izraz „zelenim sobama“ sustav će odabrati da je pridjev *zelenim* u muškome rodu, a ne u ženskome zato što *zelenim* u muškome rodu ima veću frekvenciju pojavljivanja od *zelenim* u ženskome

rodu. To pokazuje kako je kod riječi, naročito onih koji imaju mnogo mogućih gramatičkih kategorija, velika vjerojatnost za pogrešku jer najvjerojatnije gramatičke kategorije riječi nemaju veliku vjerojatnost. Da je riječ *zelenim* u, recimo, 95 % slučajeva u muškome rodu, mogla bi se napraviti pretpostavka da jest u muškome rodu. No, riječ *zelenim* u muškome je rodu samo u 53 % slučajeva po frekvencijama pojavljivanja iz baze.

Također, usluga *ispravi.me* zaprima mnogo tekstova koje treba brzo obraditi. Za spremanje leksičke baze bilo bi dobro napraviti *hash* tablicu leksičke baze jer je kompleksnost *hash* tablica O(1).

Budućem proširenju mogućnosti analizi rečenica moglo bi se pristupiti kroz statističke modele. Koristeći frekvencije pojavljivanja dostupne iz baze *hrLex* moguće je napraviti statistički model skrivenih Markovljevih lanaca za uslugu *ispravi.me*. No, uvjetna slučajna polja model je precizniji i može, ako se takvim napravi, prepoznavati nepoznate riječi za koje je prirodno da ih korisnici šalju na obradu jer u jezik dolaze nove riječi, a i baza usluge *ispravi.me* ne sadrži sve riječi hrvatskoga jezika. Prepoznavanjem nepoznatih riječi riješio bi se i problem riječi iz baze usluge *ispravi.me* koje nisu pokrivene riječima iz baze *hrLex*. Izrada uvjetnih slučajnih polja modela podrazumijevala bi izradu već spomenutih značajki prema kojima bi model zaključivao. Na temelju gramatičkih pravila trebalo bi zaključiti koje značajke postoje, odnosno koje zakonitosti u rečenici vrijede. Također, druge značajke bile bi na morfološkoj razini, poput završetaka riječi. Moguće je i razviti različite modele koji bi više odgovarali pojedinom funkcionalnom stilu hrvatskoga jezika. Dakle, koristila bi se neka vrsta strojnog učenja. Kada bi se osmislio taj sustav, kombinirala bi se analiza preko pravila i preko sustava n-grama na kojima počiva usluga strojnog provjernika pravopisa *ispravi.me*. Kombinacija pristupa prepoznata je kao dobar pristup i mnogi svjetski dobri strojni provjernici koriste i vjerojatnosne analize na temelju velikog broja podataka i analize preko jezičnih pravila [3].

Nadalje, razvijen sustav za ispravljanje gramatičkih grešaka treba povezati s uslugom *ispravi.me*. Kada bi u sustav na provjeru došla rečenica „Vidim pjesnikovz čežnja“, prvo bi se napravila pravopisna analiza koja bi pretvorila rečenicu u „Vidim pjesnikovu čežnja“. Štoviše, za ispravljanje zatipaka bi se mogla koristiti i gramatička analiza rečenice gdje bi usluga nudila ispravak zatipka s riječi koja gramatički paše u rečenici. Primjerice, za rečenicu „Vidim pjesnikovz čežnju“ sustav bi ponudio samo ispravak *pjesnikovu*, a ne *pjesnikovi*, *pjesnikove*, itd.

Baza n-grama mogla bi se koristiti za provjeru zaključaka nakon analize. Nakon što bi gramatička analiza ispravila rečenicu, rečenica bi se rastavila na trigrame, tetragrame, pentagrame te bi se provjerila frekvencija pojavljivanja tih n-grama u bazi. Ako se ti n-grami ne pojavljuju u bazi, moguće je da je došlo do greške pri gramatičkoj analizi teksta.

Zaključak

Proširenje usluge *ispravi.me* na širu i kontekstualniju analizu teksta logičan je slijed razvoja usluge. Zadatak ovog rada bio je proučiti leksičku bazu usluge *ispravi.me* i pronaći način kako bi se ona nadopunila bitnim informacijama za gramatičku analizu teksta. To je uspješno napravljeno korištenjem baze *hrLex* koja je bogata informacijama. Budući rad trebao bi potražiti veće greške u pokrivanju, proširiti listu prefiksa i završetaka kako bi se potpuno iskoristio taj pristup pokrivanju, a i razviti algoritme za raspoznavanje nepokrivenih riječi i novih riječi koje pristižu u sustav. Analiza sročnosti pridjeva i imenica u gramatičkim kategorijama pokazuje slabost takve analize u hrvatskome jeziku i s ovim tipom baze. Stoga je potrebno razviti sofisticiraniji algoritam za gramatičku analizu. Nakon ovoga može se napraviti statistički model poput skrivenih Markovljevih lanaca i uvjetovanih slučajnih polja, nekakvo strojno učenje; Taj model bi u mogućoj sinergiji s n-gramskim sustavom davao dobre rezultate. Za osmišljavanje paradigmi za analizu rečenica bilo bi dobro konzultirati se s lingvistom te osobama i tvrtkama koje u velikim količinama koriste uslugu *ispravi.me*.

Leksička baze usluge *ispravi.me* izrazito je moćan alat. N-gramski sustavi potrebni su za razvoj složenih usluga vezanih za jezik i obradu jezika (ispravi.me). Postojanjem ogromne n-gramske baze hrvatski jezik ne zaostaje mnogo za ostalim svjetskim, mnogo više korištenim jezicima. To je blago koje treba čuvati i na njemu raditi. Ovaj rad pokazao je jedan smjer kojim računalna obrada jezika može krenuti: oplemenjujući leksičku bazu usluge *ispravi.me* s gramatičkim kategorijama moguće je razviti algoritme za određivanje gramatičkih kategorija riječi u rečenici, a kada bi se spojila gramatička analiza za vjerojatnosnom koja već postoji u n-gramskom sustavu, ishodila bi vrhunska tehnologija s najboljim pristupom jezičnoj analizi.

Literatura

[1] G. Gledec, „Ispravi.me – najčešće pravopisne pogreške u 2023. godini“, fer.unizg.hr, 2024. [Online]. Dostupno: https://www.fer.unizg.hr/novosti?@=2z3tg [Datum pristupa: 7. lipnja 2024.].

[2] Nepoznat autor, „O usluzi“, ispravi.me, nepoznata godina [Online]. Dostupno: https://ispravi.me/info/ [Datum pristupa: 2. lipnja 2024.].

[3] R. Grundkiewicz, „Algorithms for Automatic Grammatical Error Correction“, doktorski rad, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznań, Poznań, 2017.

[4] D. Naber, „A rule-based style and grammar checker“, doktorski rad, Bielefeld University, Bielefeld, 2003.

[5] Nepoznat autor, „Godišnja inventura ususret 30. rođendanu“, ispravi.me, 2024. [Online]. Dostupno: https://ispravi.me/info/godisnja-inventura-ususret-30-rodendanu/ [Datum pristupa: 4. lipnja 2024.].

[6] Nepoznat autor, „gramatika“, enciklopedija.hr, nepoznata godina [Online]. Dostupno: https://enciklopedija.hr/clanak/gramatika [Datum pristupa: 11. lipnja 2024.].

[7] Nepoznat autor, „fonologija“, hjp.znanje.hr, nepoznata godina [Online]. Dostupno: https://hjp.znanje.hr/index.php?show=search_by_id&id=fFliXRI%3D&keyword=fonologija [Datum pristupa: 11. lipnja 2024.].

[8] Nepoznat autor, „morfologija“, hjp.znanje.hr, nepoznata godina [Online]. Dostupno: https://hjp.znanje.hr/index.php?show=search_by_id&id=e1hkXhE%3D&keyword=morfologija [Datum pristupa: 11. lipnja 2024.].

[9] Nepoznat autor, „sintaksa“, hjp.znanje.hr, nepoznata godina [Online]. Dostupno: https://hjp.znanje.hr/index.php?show=search_by_id&id=d19hWhY%3D&keyword=sintaksa [Datum pristupa: 11. lipnja 2024.].

[10] Nepoznat autor, „stilistika“, enciklopedija.hr, nepoznata godina [Online]. Dostupno: https://enciklopedija.hr/clanak/stilistika [Datum pristupa: 11. lipnja 2024.].

[11] J. Silić i I. Pranjković, „Gramatika hrvatskoga jezika: za gimnazije i visoka učilišta“. Zagreb: Školska knjiga, 2005.

[12] E. Barić et al., „Hrvatska gramatika“, drugo izdanje. Zagreb: Školska knjiga, 1997.

[13] Nepoznat autor, „How to Determine the Category of a Word“, oreilly.com, nepoznata godina [Online]. Dostupno: https://www.oreilly.com/library/view/natural-language-processing/9780596803346/ch05s07.html#:~:text=The%20internal%20structure%20of%20a,likely%20to%20be%20a%20noun. [Datum pristupa: 12. lipnja 2024.]

[14] Sujatha Mudadla, „What is POS Tagging“, medium.com, 2023. [Online]. Dostupno: https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186 [Datum pristupa: 13. lipnja 2024.].

[15] D. Jurafsky, J.H. Martin, „Speech and Language Processing“, web.stanford.edu, 2024. [Online]. Dostupno: https://web.stanford.edu/~jurafsky/slp3/8.pdf [Datum pristupa: 29. svibnja 2024.]

[16] Nepoznat autor, „Markovljev lanac“, enciklopedija.hr, nepoznata godina [Online]. Dostupno: https://www.enciklopedija.hr/clanak/markovljev-lanac [Datum pristupa: 13. lipnja 2024.].

[17] N. Ljubešić, „Inflectional lexicon hrLex 1.3“, clarin.si, 2019. [Online]. Dostupno: https://www.clarin.si/repository/xmlui/handle/11356/1232 [Datum pristupa: 27. svibnja 2024.].

[18] N. Ljubešić, „Inflectional lexicon hrLex 1.0“, clarin.si, 2019. [Online]. Dostupno: https://www.clarin.si/repository/xmlui/handle/11356/1056 [Datum pristupa: 27. svibnja 2024.].

[19] Nepoznat autor, „lema“, hjp.znanje.hr, nepoznata godina [Online]. Dostupno: https://hjp.znanje.hr/index.php?show=search_by_id&id=e19kUBQ%3D&keyword=lema [Datum pristupa: 12. lipnja 2024.].

[20] Nepoznat autor, „Serbo-Croatian Specifications“, nl.ijs.si, 2022. [Online]. Dostupno: https://nl.ijs.si/ME/V6/msd/html/msd-hbs.html [Datum pristupa: 29. svibnja 2024.].

Sažetak

Analizira se leksička baza usluge *ispravi.me*, sustava za računalnu provjeru pravopisa koji se desetljećima razvija na Fakultetu elektrotehnike i računarstva u Zagrebu. Leksička baza usluge temelji se na n-gramskom sustavu. Kako bi se razvila mogućnost ispravljanja ne samo pravopisnih, već i gramatičkih pogrešaka, potrebno je iz n-gramske baze stvoriti bazu koja ne sadrži samo n-grame i njihove frekvencije, nego i gramatičke kategorije riječi. Ovim je radom u velikom postotku pokriven općejezični fond riječi leksičke baze usluge uporabom javno dostupnog leksikona *hrLex* koji sadrži gramatičke kategorije riječi. Napravljen je algoritam za jednostavnu gramatičku analizu teksta te se razmatra razvoj statističkog modela za označavanje gramatičkih kategorija riječi po uzoru na skrivene Markovljeve lance i uvjetna slučajna polja (CRF).

Summary

The subject of analysis is the lexical database of the internet service *ispravi.me*, a system for spell-checking texts that is being developed at the Faculty of Electrical Engineering and Computing in Zagreb. The lexical database of the service is based on the n-gram system. In order to develop the possibility of correcting not only spelling, but also grammatical errors, it is necessary to create a database from the n-gram database that contains not only n-grams and their frequency, but also grammatical categories of words. A large portion of words from the database is covered using the public available lexicon *hrLex* which contains grammatical categories of words. An algorithm for simple grammar analysis is implemented and the development of a statistical model for marking grammatical categories of words based on hidden Markov chains and Conditional random fields in considered.