Laboratorij za tehnologije znanja (KTLab)
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Interni dokument

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 286

STROJNA PROVJERA GRAMATIKE I STILA U TEKSTOVIMA NA HRVATSKOME JEZIKU

Ognjen Lajšić

Zagreb, lipanj 2011.

Sadržaj

1. Uvod

U današnje su vrijeme prigode u kojima riječi zapisujemo izravno na papir sve rjeđe, a pri pisanju i oblikovanju tekstova sve se više služimo računalima. Računala rabimo u različitim oblicima pisane komunikacije, na njima nastaju novinski i znanstveni članci, pa čak i književnoumjetnička djela. Poneke su situacije u kojima moramo nešto napisati neformalnoga karaktera – poput neslužbenog dopisivanja među prijateljima ili na internetskim forumima – i tada je dozvoljeno slobodno se izraziti, bez potrebe da se strogo pridržavamo jezičnih normi (premda neki tu slobodu neopravdano proširuju i na opće norme ponašanja). Na internetskom forumu neslužbenoga tipa, primjerice, nije neuobičajeno niti neprihvatljivo rečenice pisati isključivo malim slovima, ne poštujući pravila o rečeničnim znakovima, ne stavljajući dijakritičke znakove tamo gdje su potrebni ili upotrebljavajući riječi koje ne pripadaju hrvatskome standardnom jeziku, poput žargonizama ili dijalektizama. Takav slobodan i neopterećen pristup pismenom izražavanju možemo dovesti u vezu s *razgovornim jezikom*,[1] što je „jezik kojim se služimo za neposredno sporazumijevanje u svakidašnjim životnim prilikama“ (Težak i Babić, 1996).

Ipak, u velikom je broju situacija uporaba razgovornog jezika neprikladna i neprihvatljiva. U bilo kakvim javnim istupima, u znanstvenim i novinskim člancima, stručnim knjigama, službenim dopisima, zakonima, životopisima i u cijelom nizu drugih pisanih vrsta od nas se očekuje da pišemo na način koji je potpuno

[1] Štoviše, iako naziv *razgovorni* upućuje na usmeno izražavanje, on po svemu odgovara i prethodno opisanom obliku pismenog izražavanja. U vrijeme kad je naziv nastao okolnosti su bile drugačije i podrazumijevalo se da sloboda koja je dozvoljena u razgovoru nije dozvoljena i u pismu (djelomično i zato što se u pisanoj komunikaciji, za razliku od govorne, mimikom i gestom nisu mogli otkloniti eventualni nesporazumi uzrokovani odstupanjem od norme). U međuvremenu, dolaskom kompjuterske i internetske revolucije ljudi su prihvatili nove vrijednosti (ponekad smislene i opravdane, ponekad ne), a pojavile su se i nove mogućnosti – pisana komunikacija sada je puno brža, lakša i neposrednija, pa je u skladu s time i lakše ispraviti eventualne nesporazume.

ili što je više moguće u skladu s gramatičkom, pravopisnom i rječničkom normom hrvatskoga jezika. U takvim situacijama koristimo se *standardnim (književnim) jezikom*, odnosno jezikom koji služi kao pouzdano sredstvo komuniciranja svih pripadnika jednoga naroda (Težak i Babić, 1996). Osim uklađenosti s jezičnim normama, zahtijeva se da se u tekstovima poštuju i odlike dobra stila, koje se prema obliku i svrsi teksta mogu razlikovati. U nekim vrstama tekstova, primjerice, dozvoljeni su subjektivni i osjećajno obojeni izrazi, dok su u drugima takvi izrazi nepoželjni, a prevladava težnja za preciznošću i objektivnošću. Postoje i izrazi koji su nepoželjni bez obzira na svrhu teksta, poput pleonazama, u kojima se suvišno gomilaju istoznačne riječi (**potencijalna mogućnost*, **no međutim*).

Još donedavno, prije pojave računala, za pronalaženje pogrešaka kojih u pisanju nismo bili svjesni ili koje su nam promaknule pri naknadnom pregledavanju zaduženi su bili isključivo lektori, ljudi s visokim stupnjem poznavanja jezika i jezičnih pitanja. Međutim, s pojavom računala i razvojem područja obrade prirodnoga jezika (engl. *natural language processing*, NLP) otvorile su se nove mogućnosti – u određenoj je mjeri mogućom postala automatska, strojna provjera jezične ispravnosti teksta. Za provjeru različitih tipova pogrešaka razvijeni su različiti programi, koji rade s manjom ili većom uspješnosti. U prvome redu tu su pravopisni provjernici (engl. *spell checkers*), ne pretjerano složeni programi koji provjeravaju jesu li sve riječi u tekstu ispravno napisane, odnosno pripadaju li leksiku nekoga jezika. Pravopisni provjernici svoj posao obavljaju prilično dobro, no uglavnom su ograničeni na pogreške koje je moguće otkriti bez sagledavanja konteksta.

Od pravopisnih su provjernika složeniji i manje uspješni gramatički i stilski provjernici (engl. *grammar and style checkers*), koji su, kao što im i samo ime govori, zaduženi za pronalaženje gramatičkih i stilskih pogrešaka. Razlog njihovoj manjoj uspješnosti jest taj što je za pronalaženje pojedinih gramatičkih i stilskih pogrešaka potrebno razumijevanje teksta na razini koja zasada za računala nije ostvariva. Štoviše, upitno je hoće li računala tu razinu ikada dostići, a sve do tog trenutka uloga koju u provjeri imaju ljudi bit će nezamjenjiva. Ipak, to ne znači da gramatički i stilski provjernici ne mogu biti korisni – usprkos lošijoj uspješnosti mogu nam poslužiti kao dodatna pomoć u pisanju, kako bismo gramatičke i stilske pogreške sveli na što je moguće manji broj. Iako se može činiti da kod izvornih govornika poznavanje gramatike nije upitno i da se gramatičke pogreške u njihovim tekstovima neće pojavljivati, istraživanja su pokazala da to ipak nije tako (Bustamante i León, 1996). Osim toga, provjernici mogu koristiti i strancima

pri učenju i lakšem savladavanju jezika, a mogu se iskoristiti i u sustavima iz područja obrade prirodnog jezika kod kojih je bitno da tekst na ulazu u sustav sadrži što manje pogrešaka.

Osnovna je motivacija nepostojanje radova niti adekvatnih alata koji se bave provjerom gramatike i stila u hrvatskome jeziku – za hrvatski jezik razvijeni su različiti pravopisni provjernici, no niti jedan gramatički i stilski. Cilj je, dakle, bio zakoračiti u to područje te razraditi i iskušati različite postupke na tekstovima na hrvatskome jeziku. U ostvarivanju gramatičkoga i stilskoga provjernika bio bi vrlo koristan sintaktički analizator, no on još nije razvijen za hrvatski jezik. Ostvareni postupci oslanjaju se stoga na statistički pristup i pristup temeljen na pravilima. Statistički postupci koji su ostvareni temelje se na različitim tipovima n-grama te na jezičnom modelu, a pristup temeljen na pravilima uporabljen je za provjeru sročnosti među imenicama i pridjevima. Na kraju su postupci eksperimentalno vrednovani i komentirana je njihova uspješnost.

Struktura ovoga rada jest sljedeća: u idućem poglavlju dana je taksonomija postupaka iz područja provjere gramatike i stila te pregled dosadašnjih radova s obzirom na iznesenu taksonomiju. U trećem poglavlju iznesena je tipologija jezičnih pogrešaka u pisanju, a različiti tipovi pogrešaka dovedeni su u vezu s vrstom provjernika u čijoj su nadležnosti. U četvrtom poglavlju detaljno su s teorijskoga i praktičnog aspekta opisani postupci razrađeni u svrhu provjere gramatike i stila u hrvatskome jeziku, a petom poglavlju izneseni su i komentirani rezultati njihova vrednovanja. U zadnjemu poglavlju dan je zaključak te su iznesene mogućnosti za budući rad. U rad je uključen i dodatak – samostalno sastavljeni korpus za vrednovanje.

2. Srodni radovi

Provjerom gramatike i stila kao temom iz područja obrade prirodnoga jezika znanstvenici se intenzivnije bave posljednjih dvadesetak godina. Iako isprva rezultati nisu bili obećavajući, danas područje provjere gramatike i stila, u skladu s razvojem različitih jezičnih alata, postaje sve perspektivnije. U ovome je poglavlju dana taksonomija postupaka iz tog područja te pregled dosadašnjih radova s obzirom na iznesenu taksonomiju. Na kraju poglavlja opisani radovi uspoređuju se s pristupom primjenjenim u ovom radu.

2.1. Taksonomija postupaka provjere gramatike

Provjera gramatičke ispravnosti tekstova složena je zadaća koja uključuje mnoge različite postupke razvijene u okviru područja obrade prirodnoga jezika. Takvi postupci neophodni su pri predobradi teksta, a mogu se razlikovati po svojim načelima rada. Ponekad nije jednostavno odrediti je li neki postupak samo alat koji služi kao puka priprema teksta za postupak provjere ili je on temeljni dio same provjere. Također, kako bi se postigli što bolji rezultati, često se kombiniraju i postupci provjere koji se razlikuju u svome pristupu. Teško je stoga postupke provjere gramatike strogo razgraničiti i odrediti jasne kriterije za određenu podjelu. Umjesto toga, podjela je stvar dogovora i možemo je donekle shvatiti kao pitanje načela rada koje dominira u postupku provjere.

Jedna moguća i uvriježena podjela (Naber, 2003; Shaalan, 2005) jest na sljedeće tri kategorije:

Provjera temeljena na pravilima – skup pravila kojim se razlikuju ispravne od neispravnih jezičnih konstrukcija primjenjuje se na tekst. Ako neki dio teksta odgovara jednome od pravila kojim se opisuje nedozvoljena jezična konstrukcija, taj je dio neispravan;

Provjera temeljena na sintaktičkoj analizi – čitav se tekst sintaktički ana-

lizira (engl. *syntactic analysis, parsing*), odnosno svakoj se rečenici pokušava odrediti gramatička struktura s obzirom na određenu gramatiku (skup pravila kojim se određuje ispravan ustroj rečenice). Ako sintaktička analiza neke rečenice ne uspije, rečenica je gramatički neispravna;

Provjera temeljena na statistici – na velikom korpusu gramatički ispravnog teksta utvrđuje se koji su slijedovi riječi ili oznaka vrste riječi vjerojatniji, a koji su manje vjerojatni. Ta informacija primjenjuje se kasnije pri provjeri gramatike.

Osim te podjele, u literaturi se spominje i podjela u kojoj je izostavljena sintaktička analiza kao moguća osnova postupka provjere (Henrich i Reuter, 2009). Nije uvijek i u potpunosti jasno treba li onda postupke koji uključuju sintaktičku analizu svrstati u provjere temeljene na pravilima ili one temeljene na statistici – to ovisi o značajkama samoga parsera, ali i o točki gledišta, prema kojoj način na koji je parser nastao (učenjem na korpusu ili ručnim definiranjem pravila) može biti više ili manje važan u odnosu na, primjerice, način na koji se otkriva točan položaj i tip pogreške. U ovome radu rabit će se podjela koja uključuje kategoriju sintaktičke analize kao osnove u postupku provjere.

2.2. Postupci temeljeni na pravilima

Kod postupaka temeljenih na pravilima u prvome je planu znanje o neispravnim jezičnim konstrukcijama. Pri izradi takvog postupka pokušava se uočiti što veći broj slučajeva u kojima se često griješi ili je moguće pogriješiti, i takvi se slučajevi zatim opisuju pravilima. Primjerice, u hrvatskom jeziku često se rabi nepravilna konstrukcija *no međutim*. Pravilom se može definirati da se na takvu konstrukciju upozori svaki put kada se ona pojavi u tekstu te da se predlože mogući ispravni oblici. Pravila se ne moraju odnositi samo na konkretne riječi, moguće ih je i apstrahirati na metajezične entitete poput oznaka vrste riječi ili sintaktičkih skupina (za to su nužni odgovarajući alati – označivač vrsta riječi odnosno plitki parser). Primjer za to jest pravilo kojim se definira da u engleskom jeziku dva člana (engl. *article*) ne smiju slijediti jedan iza drugoga.

Problem je kod pristupa temeljenog na pravilima što mogućnosti za pogrešku ima previše da bismo ih sve mogli predvidjeti. Čak i najiscrpniji i najopsežniji sustavi naići će na pogreške koje nisu obuhvaćene niti jednim njihovim pravilom. Osim toga, pristup temeljen na pravilima nije jezično nezavisan; pravila izrađena

za jedan jezik ne mogu se primijeniti na drugi jezik, već se moraju za svaki jezik izraditi zasebno. S druge strane, takav pristup ima i svoje prednosti. Na problematične jezične konstrukcije upozorava se izravno i uz to je moguće ponuditi detaljno obrazloženje problema, prijedloge ispravnih konstrukcija, pa čak i pojašnjenja gramatičkih pravila. Sustav je moguće graditi inkrementalno, neprestano dodavati nova pravila i tako ga poboljšavati, a jednostavno ga je i podesiti i onemogućiti primjenu pravila koja u određenom kontekstu nisu nužna ili korisna.

Postupci temeljeni na pravilima u svojoj se osnovi međusobno ne razlikuju mnogo. Svi se koriste označivačem vrsta riječi, a neki tome pridodaju i morfološki analizator te plitki parser. Te jezične alate primjenjuju u predobradi teksta, a informacijom o tekstu koju im oni pružaju kasnije se koriste pri definiranju pravila. Ono u čemu se postupci ipak mogu razlikovati jest način na koji su takvi jezični alati izgrađeni. Primjerice, označivač vrsta riječi može i sam biti temeljen na pravilima, kao u (Gill i Lehal, 2008), ili pak može nastati treniranjem na korpusu, tj. statističkim pristupom (Carlberger et al., 2002; Naber, 2003). Ovdje je vidljiva i proizvoljnost u smještanju postupka provjere u određenu kategoriju s obzirom na njegova načela rada – ako na označivač temeljen na statistici gledamo kao na temeljni dio same provjere, postupak provjere svrstat ćemo među hibridne. Neki su autori skloni takvom poimanju (Carlberger et al., 2002), no mi ćemo se držati gledišta prema kojem je svaki jezični alat osim sintaktičkog analizatora samo alat za predobradu, pa ćemo i takve postupke smjestiti među one temeljene na pravilima.

Postupci iz ove domene često su primjenjivani u praksi i postižu relativno dobre rezultate. Najčešće su nešto bolje preciznosti u odnosu na odziv, što možemo objasniti time da će dobro definirano pravilo rijetko ili nikada neku ispravnu konstrukciju prepoznati kao neispravnu, dok je s druge strane pravilima nemoguće pokriti sve moguće jezične pogreške. Javno dostupan sustav otvorenoga koda za provjeru gramatike i stila napravio je Naber (2003). Isprva su implementirana samo pravila za engleski, no omogućena je i njihova jednostavna izrada u XML-datotekama, pa su u međuvremenu u većoj ili manjoj mjeri dodana pravila za brojne druge jezike (za hrvatski još nije dodano niti jedno pravilo).[1] U okviru provjere temeljene na pravilima valja spomenuti još i provjernike rađene za švedski (Carlberger et al., 2002; Eeg-Olofsson i Knutsson, 2003), brazilski portugalski (Kinoshita et al., 2006), njemački (Schmidt-Wigger, 1998), punjabi (Gill i Lehal,

[1] `http://www.languagetool.org/languages`

2008) i korejski (Young-Soog, 1998). U tim radovima, izuzevši onaj za njemački, fokus je prvenstveno na provjeri gramatike, dok se provjera stila spominje samo usputno.

2.3. Postupci temeljeni na sintaktičkoj analizi

Sintaktičkom analizom pokušava se utvrditi gramatička struktura rečenice s obzirom na određenu gramatiku. Ako sintaktička analiza ne uspije, rečenica nije u skladu s tom gramatikom, odnosno gramatički je neispravna. Primjenom parsera neposredno se dakle donosi sud o ispravnosti rečenice, pa na njega gledamo kao na temeljni dio provjere. Stoga mu se, iako je i sam svojevrstan jezični alat, uloga u postupku provjere razlikuje od one ostalih jezičnih alata te ga izdvajamo u zasebnu kategoriju. Međutim, uporabom parsera možemo samo utvrditi je li rečenica ispravna ili nije (u odnosu na gramatiku nad kojom je parser izgrađen). Kako bismo saznali točan položaj i tip pogreške ili predložili ispravne mogućnosti, moramo se poslužiti drugim metodama (Hein, 1998), koje mogu uključivati i postupak temeljen na pravilima. Tako će, ako pod provjerom podrazumijevamo i ostale stupnjeve, a ne samo otkrivanje pogreške, postupci koji uključuju sintaktičku analizu često biti hibridnoga tipa.

Zasada su u praksi razvijene metode kojima se utvrđuje točan položaj i tip pogreške temeljene na dvama različitim pristupima: *predviđanje pogrešaka* (engl. *error anticipation*) i *popuštanje uvjeta* (engl. *constraint relaxation*). Pristup s predviđanjem pogrešaka odnosi se na znanje o mogućim pogreškama i sličan je ili u potpunosti odgovara pristupu temeljenom na pravilima. Jedna mogućnost jest proširenje gramatike tako da prihvaća i pogrešne, gramatički neispravne strukture. Rečenicama koje su prihvaćene samo uz produkciju kojom se opisuje neispravna struktura može se tako odmah postaviti ispravna dijagnoza. Gramatike se mogu u potpunosti razdijeliti na one kojima se opisuju ispravne i one kojima se opisuju neispravne rečenične strukture, i u tom smislu razlikujemo pozitivnu i negativnu gramatiku (Clément et al., 2011). Druga je mogućnost kod pristupa s predviđanjem pogrešaka kombiniranje pozitivne gramatike i pravila – pozitivna gramatika iskoristi se kako bi se otkrile neispravne rečenice i zatim se samo na te rečenice primjenjuju pravila kako bi se eventualno otkrio položaj i tip pogreške; na taj način smanjuje se broj lažnih uzbuna koje uzrokuje nepreciznost pravila.

Kod pristupa temeljenog na popuštanju uvjeta ciljano se umanjuju uvjeti potrebni da bi parser prihvatio rečenicu kao ispravnu. Ako znamo da je parsanje

uspjelo tek nakon što smo uklonili određeni uvjet, onda znamo da upravo taj uvjet nije zadovoljen u rečenici. Ovdje razlikujemo strukturne i nestrukturne pogreške. Prve se odnose na pogreške kojima je promijenjena struktura rečenice, a izazvane su umetanjem i brisanjem riječi ili promjenom redoslijeda riječi. Drugi su tip pogrešaka one koje ne utječu na strukturu rečenice, npr. slaganje riječi u rodu, broju ili padežu. Popuštanje uvjeta uobičajeno se odnosi na na nestrukturne pogreške.

Prednost pristupa temeljenog na sintaktičkoj analizi jest u tome što najbolje odgovara samome problemu – nas zanima pripadaju li sve konstrukcije u tekstu gramatici nekoga jezika, i na to pitanje parser nam daje izravan odgovor. Ako je gramatika jezika potpuna, odnosno njome su obuhvaćene sve moguće ispravne jezične strukture, onda će i sve moguće pogreške biti otkrivene. No, problem je što su prirodni jezici vrlo složeni i za mnoge od njih zasada, usprkos mnogim gramatičkim teorijama, nisu razvijeni parseri koji parsaju s obzirom na potpunu gramatiku jezika. Tako se neuspjeh pri parsanju neke rečenice ipak ne može sa sigurnošću pripisati gramatičkoj pogrešci, s obzirom na to da može proizlaziti iz nepotpunosti gramatike. Sintaktička analiza pritom je sve kompliciranija što jezik više naginje k sintetičnim jezicima (kao što je slučaj i s hrvatskim), a manje je problematična na strani analitičkih jezika.[2] Postupci temeljeni na sintaktičkoj analizi ipak su u prednosti pred ostalim dvama postupcima u tome što sagledavaju čitavu rečenicu odjednom i tako ne gube nikakvu sintaktičku informaciju.

2.4. Postupci temeljeni na statistici

Kod postupaka temeljenih na statistici iskorištava se statistička informacija o učestalosti ili vjerojatnosti pojavljivanja jezičnih konstrukcija. Što se neka konstrukcija više puta pojavila u korpusu, to je vjerojatnije da će i u nekom drugom kontekstu biti ispravna. Pri provjeri postavlja se prag i na sve se konstrukcije čija je učestalost ili vjerojatnost pojavljivanja manja od tog praga upozorava kao na potencijalno neispravne. Prednost takvoga pristupa jest njegova jednostavnost i laka izvedivost. Nedostaci su mu orijentiranost na manje konstrukcije umjesto na čitave rečenice, nemogućnost postavljanja točne dijagnoze pogreške i davanja objašnjenja korisniku te nepreciznost, često uzrokovana i, u gramatičkom smislu,

[2]Sintetični jezici nastoje ujediniti više morfema u jednoj fonetskoj riječi; u prenošenju značenja oslanjaju se na fleksiju, a poredak riječi manje je bitan. U suprotnosti su s njima analitički jezici, u kojima je značenje sadržano u poretku riječi, dok je fleksija minimalna.

nedovoljno kvalitetnim tekstom u korpusu. Također, ideja proizvoljnog praga kao okosnice pri odluci o ispravnosti konstrukcije ne podudara se s idejom da su rečenice uobičajeno ili točne ili netočne. Budući da je ovaj pristup jedan od pristupa na kojima se zasniva provjera gramatike opisana u ovome radu, radovi iz tog područja detaljnije su prikazani u nastavku.

2.4.1. Prvi pokušaji

Kernick i Powers (1996) analiziraju različite pristupe provjeri gramatike. Zaključuju da statistički pristup najbolje odgovara problemu, ali da provjeri trebamo pristupati pitajući se za koju je od mogućih rečenica najvjerojatnije da je ispravna, a ne je li rečenica koju provjeravamo ispravna. To potkrepljuju primjerom dviju riječi vrlo čestih u engleskome jeziku, *like* i *down*, koje po svojoj vrsti pripadaju u čak sedam kategorija. Budući da takvih riječi u engleskome ima puno, ispravnost rečenice ne može se odrediti promatrajući rečenicu samu za sebe; umjesto toga, potrebno je pronaći moguće varijacije rečenice i odrediti koja je od njih najvjerojatnija. Rješenje problema temelje na konceptu takozvane diferencijalne gramatike. Cilj je diferencijalne gramatike pronaći skup minimalnih konteksta na temelju kojih je moguće razlikovati ispravne od krivo upotrijebljenih riječi. Taj koncept primjenjuju na riječima koje se često zamjenjuju u engleskome jeziku (primjer su riječi *form* i *from*). Sve statističke informacije prikupljaju iz korpusa koji sadrži više od 100 milijuna riječi, a sastoji se uglavnom od novinskih članaka. Nedetaljno testiranje na malom broju rečenica pokazalo je ohrabrujuće rezultate: otkrivene su sve netočne rečenice, a u samo 1.9% slučajeva točne su rečenice krivo označene kao netočne.

Suprotno tezi Kernicka i Powersa da je bolje provjeravati koja je od mogućih varijanti rečenice najvjerojatnija, Alam et al. (2006) pogrešne rečenice pokušavaju otkriti jezičnim modelom, odnosno koristeći se isključivo informacijom o vjerojatnostima trigrama u rečenici. Razmatraju mogućnost uporabe trigrama sačinjenih od riječi, ali se zbog nedostatne veličine korpusa koji imaju na raspolaganju odlučuju za trigrame sačinjene od oznaka vrste riječi. Pretpostavka je da je rečenica točna ako je umnožak vjerojatnosti njezinih trigrama veći od određenog praga. Prag je u ovome slučaju postavljen na 0, što znači da je točna svaka rečenica u kojoj su se svi trigrami barem jednom pojavili u korpusu. Niti jedna metoda zaglađivanja nije primijenjena uz jezični model – ako se samo jedan od trigrama u rečenici nije pojavio u korpusu, rečenica se smatra netočnom. Koliko

je moguće iščitati iz rada, postupak je evaluiran isključivo na točnim rečenicama. Provjernik je ispravno označio 63% rečenica na engleskome (od 866 rečenica, 545 je ispravno prepoznao kao točne) i 53.7% na bengalskome (203 ispravno prepoznate rečenice od njih 378). Ti rezultati dobiveni su uz vlastoručno označavanje vrste riječi; za očekivati je da je stvarna uspješnost provjernika lošija, s obzirom na to da automatsko označavanje vrste riječi nikad nije u potpunosti precizno. Također, za ozbiljnije prosudbe o uspješnosti nedostaju rezultati provjernika u otkrivanju netočnih rečenica.

Opisani radovi načeli su temu rješavanja problema provjere gramatike statističkim pristupom. Iako su navedene metode zanimljive kao izvor ideja i inspiracije, nedovoljno su razrađene da bi poslužile kao samostalna osnova provjernicima gramatike. Postupak Kernicka i Powersa polučio je dobre rezultate – doduše, to možemo reći samo na temelju preliminarnoga ispitivanja – no mana mu je što je orijentiran na pojedinačne, specifične slučajeve gramatičkih netočnosti. Kako bi takav postupak bio zbilja uspješan, potrebno je identificirati što veći broj slučajeva u kojima ljudi griješe, što zahtijeva detaljnu jezičnu analizu, a čak i uz takvu analizu nije zajamčeno da se neće pojaviti dotad nezamijećene pogreške. Također, problem je i što takav pristup nije univerzalan, s obzirom na to da u različitim jezicima postoje različite pogreške, pa je za svaki jezik potrebno iznova raditi analizu.

Alam i suradnici krenuli su u drugom, općenitijem smjeru. Umjesto onoga što bi *moglo biti krivo*, oni se fokusiraju na ono za što su *sigurni da je točno*, a krivim proglašavaju sve što se razlikuje od toga. Međutim, problem u njihovu pristupu proizlazi iz nedovoljne razrađenosti metode i iz premala korpusa. Prirodni je jezik vrlo složen sustav te ga nije moguće obuhvatiti samo jednostavnim jezičnim modelom poput njihova, izgrađenim nad nevelikim korpusom. Tako je velik broj jezičnih konstrukcija izostavljen čak i nakon poopćenja jezičnog modela s trigrama sačinjenih od riječi na trigrame sačinjene od oznaka vrste riječi. Zbog toga se, djelomice i zato što na jezičnom modelu nije primijenjeno zaglađivanje, javlja velik broj lažnih uzbuna (engl. *false alarms*), odnosno situacija kada je točna rečenica neispravno označena kao netočna. S druge strane, poopćenje jezičnog modela donosi i smanjenje sigurnosti u ono što je točno, s obzirom na to da je podjela na vrste riječi pregruba da bismo njome uvijek mogli razlučiti između ispravnih i neispravnih sintagmi. Primjer za to su sintagme *I am nice* i *You am nice* – budući da su označene istim slijedom oznaka vrste riječi, obje će sintagme biti prepoznate na jednak način, iako je prva ispravna, a druga neispravna. Tome

treba pribrojiti i to da su trigramskim modelom obuhvaćene samo sintaktičke veze među trima riječima u nizu, iako riječi mogu biti sintaktički povezane i na puno većim „udaljenostima“. Tako će značajan broj pogrešaka proći nezamijećeno, odnosno netočne će rečenice biti neispravno prepoznate kao točne.

2.4.2. LISGrammarChecker

Henrich i Reuter (2009) nadišli su razinu svojih prethodnika i osmislili zasada najkompletnije rješenje temeljeno na statistici. Njihov program naziva je LISGrammarChecker, što je skraćeno od *Language Independent Statistical Grammar Checker*. Za takav, jezično nezavisan statistički pristup odlučili su se jer su svi uspješni provjernici gramatike do tada bili temeljeni na pravilima. Pristup temeljen na pravilima može biti efikasan, no njegov je izrazit problem jezična zavisnost. Pravila se za svaki jezik moraju izraditi zasebno, a neki su jezici suviše složeni ili ih govori premali broj ljudi da bi izgradnja pravila za njih bila isplativa. Za razliku od sustava temeljenih na pravilima, LISGrammarChecker ne zahtijeva nikakvo lingvističko znanje, a sve informacije potrebne za rad crpi iz statističkih podataka. Njegov princip rada detaljnije je objašnjen u nastavku.

LISGrammarChecker provjeri gramatike paralelno pristupa s dvaju različitih aspekata. Jedan je općenit i u njemu se provjerava svako odstupanje od onoga što se smatra ispravnim na temelju statističkih podataka, pri čemu se oslanja na n-grame sačinjene od riječi i od oznaka vrste riječi. U drugome je naglasak na morfosintaktičkom značajkama teksta, odnosno na specifičnim gramatičkim problemima. Tako prvi aspekt provjere možemo usporediti s onim Alama i suradnika, a drugi s radom Kernicka i Powersa (iako je i dalje općenitiji od njihova, s obzirom na to da se fokusira na poklapanje u pojedinim morfosintaktičkim kategorijama, a ne na točno određene i konkretne riječi). Rad programa podijeljen je u dvije faze. U prvoj fazi program prikuplja informacije iz velike količine statističkih podataka, odnosno iz korpusa u kojemu su prethodno označene vrste riječi, dok je u drugoj fazi moguće na temelju prikupljenih informacija provjeriti gramatičku ispravnost nekog teksta.

U općenitu provjeru gramatičke ispravnosti, onu s n-gramima, uključen je velik broj različitih tipova n-grama, kako bi se što bolje obuhvatile složene jezične strukture. U obzir se uzimaju n-grami sačinjeni od dviju pa sve do pet riječi (bigrami, trigrami, kvadrigrami, kvintagrami) te, jednako tako, od dviju do pet oznaka vrste riječi. Uz to, kao posebna kategorija sagledavaju se i oznake vrste

svih riječi u rečenici, odnosno n-grami koji se sastoje od svih oznaka u jednoj rečenici, ma koliko duga ona bila. Na kraju, u provjeru ulaze i hibridni n-grami, sačinjeni jednim dijelom od riječi, a drugim od oznaka vrste riječi (konkretno, bigrami s jednom riječi i jednom oznakom i trigrami s riječi u sredini i oznakama sa svake strane). Prilikom treniranja programa korpus se analizira i razlaže na svaki od spomenutih tipova n-grama, te se svi n-grami koji se pojavljuju u korpusu trajno pohranjuju u bazu podataka. Tako pohranjeni podaci potom se rabe pri provjeri gramatike.

Provjera gramatičke ispravnosti nekog teksta temelji se na *pretpostavkama pogreške* (engl. *error assumptions*). Svaki n-gram koji se nije pojavio u tekstu koji provjeravamo, a pojavio se u korpusu na kojemu je program treniran, zapravo je jedna pretpostavka pogreške. Svaka pretpostavka pogreške nosi određenu težinu s obzirom na pripadajući tip n-grama. Ako se unutar jedne rečenice pojavi dovoljan broj takvih pretpostavki, rečenica se smatra pogrešnom. Sam postupak provjere započinje na sličan način kao i obrada korpusa pri treniranju programa. Tekst se najprije odvaja na rečenice i riječi, a riječi se analiziraju tako da im se dodaju oznake njihove vrste. Provjera se zatim odvija za svaku rečenicu zasebno. U odgovarajućoj tablici u bazi podataka provjerava se postoji li n-gram koji se sastoji od oznaka svih riječi u čitavoj rečenici. Ako ne postoji, to je onda jedna pretpostavka pogreške, i postupak se u tom slučaju nastavlja sa slijedom od pet oznaka. Ako se neki od 5-grama oznaka nije pojavio u korpusu, bilježi se njegova pozicija i unutar njega se nastavlja s provjerom slijedova od četiri oznake, odnosno 4-gramima. Postupkom se tako tamo gdje je potrebno provjeravaju sve manji n-grami, sve do bigrama. Na sličan način odvija se i provjera hibridnih n-grama te n-grama sačinjenih od riječi, uz napomenu da je za potonje omogućena i uporaba internetskih tražilica (ako se n-gram iz teksta nije pojavio u bazi, sljedeći je korak provjera njegove učestalosti na Internetu, i u ovisnosti o tom ishodu nastavlja se s provjerom manjih n-grama). Na kraju se sve pretpostavke pogreške zbrajaju, pri čemu se u obzir uzima značaj koji imaju za provjeru ispravnosti. Za provjeru je, primjerice, nepojavljivanje bigrama sačinjenog od oznaka u bazi podataka nešto značajnije od nepojavljivanja bigrama sačinjenog od riječi, a mnogo značajnije od nepojavljivanja 5-grama sačinjenog od riječi. Zbroj pretpostavki uspoređuje se s unaprijed definiranim pragom i u skladu s tim donosi se sud o ispravnosti rečenice.

U rečenicama koje su proglašene pogrešnima određuje se koje su riječi potencijalno uzrokovale takvo stanje. Za te riječi predlažu se najvjerojatnije zamjene

s obzirom na susjedne riječi. Najprije se u bazi traži 5-gram s potencijalno pogrešnom riječi u sredini, pri čemu se umjesto te riječi stavlja znak zvjezdice, koji označava da tu može stajati bilo koja riječ. Ako takav upit ne da nikakve rezultate, pokušava se ista stvar s trigramom. Iz mogućih zamjena na kraju treba odabrati najbolju, odnosno onu koja će biti predložena, što je moguće napraviti na dva načina. Prvi je da se odabere riječ iz sintagme koja se najviše puta pojavila u korpusu (u tu svrhu u bazi se pohranjuje i informacija o broju pojavljivanja pojedinih n-grama u korpusu), a drugi je da se odabere riječ koja je potencijalno pogrešnoj riječi najsličnija.

Uz prethodno opisanu, općenitu provjeru gramatičke ispravnosti LISGrammarChecker omogućuje i provjeravanje dvaju specifičnih gramatičkih problema – slaganje vremenskog priloga s vremenom pripadajućeg mu glagola te slaganje pridjeva i pripadajuće mu imenice u rodu, broju i padežu. Ideja je slična onoj za općenitu provjeru i svodi se ponovno na n-grame riječi i oznaka. Za prvi problem pohranjuju se u bazu svi vremenski prilozi (npr. *yesterday*) pronađeni u korpusu u paru s oznakom vremena pripadajućih glagola – za glagol *stayed*, primjerice, pamti se oznaka `verb (past tense)`. Za drugi problem pohranjuju se pridjevi u paru s imenicama koje ti pridjevi opisuju. Te informacije rabe se zatim pri provjeri gramatike svaki put kad se u nekoj rečenici zajedno pojave oznake vremenskog priloga i glagola odnosno, u drugom slučaju, pridjeva i imenice. Ako kombinacija koja se pojavila u tekstu ne postoji u bazi podataka, smatra se netočnom i pretpostavka pogreške pribraja se ukupnoj sumi.

2.5. Pristup primjenjen u ovome radu

U sklopu ovoga rada razrađeni su i iskušani postupci temeljeni na jezičnom modelu izgrađenom nad oznakama, poput onoga u (Alam et al., 2006), te postupci temeljeni na različitim tipovima n-grama, poput onih u (Henrich i Reuter, 2009). Međutim, postupak s jezičnim modelom razlikuje se od spomenutog postupka u tome što se uz njega primjenjuje zaglađivanje, dok se n-grami primjenjuju zasebno i što je za primjenu svakoga tipa n-grama provedeno detaljno ispitivanje, za razliku od (Henrich i Reuter, 2009). Uz ove postupak, implementiran je i postupak provjeru sročnosti među imenicama i pridjevima temeljen na pravilima.

3. Tipologija jezičnih pogrešaka u pisanju

U ovome poglavlju dana je tipologija jezičnih pogrešaka u pisanju, najprije kroz različite lingvističke kategorije, a zatim kroz podjelu koja je više računarski orijentirana.

3.1. Tipologija s obzirom na lingvističke kategorije

Tipove jezičnih pogrešaka u hrvatskom jeziku razmotrit ćemo kroz sljedeće lingvističke kategorije: pravopis, fonologija, morfologija, sintaksa, semantika i stil. Te tipove pogrešaka dovest ćemo u vezu s provjernicima pravopisa te provjernicima gramatike i stila, opisujući koji je tip u domeni kojega od ovih dvaju provjernika i zašto.

3.1.1. Pravopis

Pravopis je skup pravila koja određuju na koji ćemo način pri pisanju kojega jezika upotrebljavati sve pismene znakove, a to su slova te rečenični i pravopisni znakovi, uključujući bjeline (Težak i Babić, 1996). Iako su pravopis i gramatika dvije odvojene cjeline, one se u mnogome dodiruju i preklapaju. Jedan od glavnih dijelova u pravopisnim priručnicima, pisanje glasova i glasovnih skupova poput *č*, *ć*, *dž*, *đ*, *h*, *ije-je-e-i*, prije je područje gramatike (konkretno, pravogovora, fonetike i fonologije) nego pravopisa. I pravila o rečeničnim znakovima djelomično su vezanu uz gramatiku, s obzirom na to da utječu na vrstu i strukturu rečenice, što je predmet proučavanja sintakse.

Pravopisne pogreške deklarativno su u domeni pravopisnih provjernika, no u praksi to nije u potpunosti tako. Pravopisni provjernici uobičajeno su izvedeni

tako da svaku riječ iz teksta provjeravaju zasebno, ne uzimajući u obzir kontekst. Budući da su mnoga pravopisna pravila uvjetovana kontekstom, što je vidljivo i iz dodirnih točaka pravopisa i gramatike, jasno je da pravopisni provjernici ne obavljaju svoju zadaću u cijelosti u skladu sa svojim nazivom. Takva pravopisna pravila, koja se oslanjaju na kontekst, mogla bi se uspješno uklopiti u provjernike gramatike i stila. To su, npr., pravila o pisanju velikog i malog slova (**Europska Unija* > *Europska unija*; veliko slovo na početku rečenice) te rečeničnih znakova (zarez ispred suprotnih veznika), ili pak pravilo o pisanju futura prvog (**biti ću* > *bit ću*) i prijedloga *s/sa*.

3.1.2. Fonologija

Fonologija ili glasoslovlje jezikoslovna je disciplina koja proučava jezičnu ulogu i ponašanje osnovnih govornih jedinica, fonema. Prema modernom tumačenju, fonologija je dio gramatike, uz fonetiku, morfologiju, sintaksu, semantiku i pragmatiku. Primjeri pogrešaka u domeni fonologije jesu izostavljanje kojeg slova pri pisanju riječi stranoga podrijetla (**delikvent*, **intezivan*, **transplatacija* > *delinkvent*, *intenzivan*, *transplantacija*) te neprovođenje ili neispravno provođenje glasovnih promjena (**jedamput*, **stanbeni*, **potčiniti*, **izšarati*, **najači* > *jedanput*, *stambeni*, *podčiniti*, *išarati*, *najjači*). S obzirom na to da takvim pogreškama nastaju oblici koji nisu dijelom hrvatskoga jezika, fonologija je, iako gramatička disciplina, u potpunosti u nadležnosti pravopisnih provjernika.

3.1.3. Morfologija

Morfologija ili oblikoslovlje grana je jezikoslovlja koja proučava sustav jezičnih oblika te načine na koje se riječi u nekom jeziku oblikuju i mijenjaju. Dva su osnovna načina promjene riječi, fleksija i derivacija. Prvim se načinom tvore različiti oblici jednog te istog leksema, dok se drugim iz jednoga leksema tvore novi, po značenju bliski leksemi. Jedan oblik fleksije u hrvatskom jeziku jest deklinacija ili sklonidba, odnosno promjena riječi po padežima. Tako je jedna moguća pogreška morfološkoga tipa neispravno sklanjanje neke riječi (npr., akuzativ zamjenice *svi* jednak je nominativu, a ne genitivu *svih*, kako se često griješi; zatim, pravilo je da se posvojni pridjevi, tj. oni koji završavaju na *-ov*, *-ev* i *-in*, sklanjaju samo po imeničnoj sklonidbi, dok se odnosni pridjevi, tj. oni koji završavaju na *-ski/-ški/-čki*, *-nji*, *-ni*, *-ji*, sklanjaju samo po pridjevnoj sklonidbi). Kako je za otkrivanje ovakvih pogrešaka potrebno znanje o kontekstu, morfologija i s prak-

tičnog i s teoretskog aspekta pripada provjernicima gramatike i stila. Pritom treba napomenuti da neka morfološka pravila, poput onih o sklonidbi posvojnih i odnosnih pridjeva, zalaze u pitanje stila – naime, ta pravila vrijede samo za visoki i formalni stil, dok su u slobodnijim stilovima odstupanja od njih dopuštena.

3.1.4. Sintaksa

Sintaksa ili skladnja dio je gramatike koji proučava razmještaj i međusobno slaganje riječi u rečeničnim dijelovima (sintaksa izraza) te poredak i službu riječi i rečeničnih dijelova u rečenici (sintaksa rečenice). U hrvatskome jeziku poredak je rečeničnih dijelova (u ovom kontekstu: riječi ili sintagme koje u rečenici imaju neku službu, kao što su subjekt, predikat, objekt) slobodan (Težak i Babić, 1996). To možemo vidjeti na primjeru sljedećih rečenica:

1. *Čovjek pogledaše ženu.*
2. *Čovjek ženu pogledaše.*
3. *Ženu pogledaše čovjek.*
4. *Ženu čovjek pogledaše.*
5. *Pogledaše čovjek ženu.*
6. *Pogledaše ženu čovjek.*

Iz tih je rečenica vidljivo da se subjekt (*čovjek*), predikat (*pogledaše*) i objekt (*ženu*) mogu poredati na bilo koji način; rečenice su ispravne neovisno o njihovu položaju. Međutim, proizvoljnost pri razmještanju riječi unutar pojedinih rečeničnih dijelova ograničena je ili je uopće nema. Tako rečenicu *Pogledat ćemo film*, u kojoj je *Pogledat ćemo* predikatska sintagma, možemo urediti samo na četiri načina. Od toga su dva izrazito pjesnički obilježena i ne pripadaju formalnom stilu (*Film pogledat ćemo*, *Pogledati film ćemo*). Rečenice koje počinju s *ćemo* nisu ni u kojem slučaju ispravne zbog pravila koje kaže da zanaglasnica ne smije doći na prvo mjesto u rečenici. Uz red riječi u rečenici, sintaksa se bavi i sročnošću, odnosno slaganjem riječi po rodu, broju, padežu i licu. Možemo zaključiti da se sintaktičke pogreške odnose na nepravilnosti u položaju riječi u rečenici ili u njihovoj sročnosti. Takve pogreške u izravnoj su nadležnosti provjernika gramatike i stila.

3.1.5. Semantika

Semantika je jezikoslovna disciplina koja se bavi proučavanjem riječi kao nosilaca značenja, kao sredstva za označivanje predmeta, pojava i odnosa u materijalnom i duhovnom svijetu. U semantici su važni pojmovi koji govore o značenjskim odnosima među riječima: homonimi, paronimi, sinonimi, antonimi, itd. Najčešće su semantičke pogreške zamjene paronima, odnosno dvaju riječi koje se slično pišu, slično zvuče ili su bliske po značenju; npr., često dolazi do zamjena paronima poput riječi *blagdan* i *praznik*, *elektronski* i *elektronički*, *tok* i *tijek*, *desetci* i *desetine*. I pleonazmi kao stilske pogreške imaju semantička obilježja, s obzirom na to da se radi o suvišnom gomilanju *istoznačnih* riječi. Semantika nije bila tradicionalni dio gramatike (to su bile samo morfologija i sintaksa), no po novijim podjelama uvrštava se u nju. Po toj podjeli, a i zbog činjenice da odluka o pogrešnoj uporabi riječi s obzirom na njezino značenje ima smisla samo ako se sagleda kontekst, semantičke su pogreške u nadležnosti provjere gramatike i stila.

3.1.6. Stil

Stil je, kratko i jasno, način usmenog ili pismenog izražavanja (Težak i Babić, 1996). Njime se bavi posebna jezičnoznanstvena grana, stilistika, proučavajući izražajne vrijednosti jezičnih sredstava i učinak koji je postigao izabrani izraz. Kao osnovne odlike dobra stila obično se navode jasnoća, istinitost (točnost) i ljepota. Iz toga se izvode i neke druge osobine, kao što su jezgrovitost, jednostavnost, logičnost, muzikalnost, ritmičnost, slikovitost itd. Međutim, ta se obilježja u tekstovima pisanima s različitom svrhom ne odražavaju na isti način – lijepo u romanu ne mora biti lijepo i u znanstvenoj raspravi, istinito u bajci ne mora biti istinito i u novinskoj vijesti, jasno u pjesmi ne mora biti jasno i u službenom dopisu.

S obzirom na neke osobitosti u tekstovima određene namjene i oblika, možemo govoriti o više vrsta stilova. U hrvatskome standardnom jeziku razlikujemo pet funkcionalnih stilova: znanstveni, administrativni, popularnoznanstveni, novinski i književnoumjetnički stil. Ti se stilovi mogu grubo podijeliti na racionalne i emocionalne, s obzirom na to je li im svrha da djeluju prvenstveno na razum ili na emocije. Tako su znanstveni i administrativni stil strogo racionalni, popularnoznanstveni i novinski pomiruju i sjedinjuju osobine jednoga i drugoga, a književnoumjetnički stil prvenstveno je emocionalan. Racionalni stilovi manje su slobodni od emocionalnih i u njih je veća potreba za formalnim izražavanjem. U

znanstvenom stilu, koji je ujedno i najstroži funkcionalni stil, zahtijeva se potpuna usklađenost s gramatičkom, pravopisnom i rječničkom normom hrvatskoga jezika te izbjegavanje subjektivnosti, dok je s druge strane u književnoumjetničkom stilu dozvoljeno gotovo bilo što (naravno, ako ima svrhu i umjetničko opravdanje).

U skladu s tim ima smisla govoriti o strojnoj provjeri stila samo kod formalnijih, racionalnih stilova (umjetničko opravdanje kao koncept još nije u domeni računala). I u tom će slučaju provjera imati tim više smisla što je više specijalizirana, odnosno usmjerena na jedno usko stilsko područje, jer se od područja do područja mogu razlikovati vokabular, izgradnja rečenice i sl. Ipak, neka su izražajna sredstva nedobrodošla gotovo u svim tekstovima formalna karaktera, poput pleonazama, dijalektizama, žargonizama, arhaizama itd.

3.2. Strukturne i nestrukturne pogreške

Još je jedna korisna podjela pogrešaka po tipu ona na strukturne i nestrukturne pogreške. Podjela je, za razliku od podjele u prethodnom poglavlju, više računarski orijentirana, a proizašla je iz različitog utjecaja pogrešaka na sintaktičku analizu. Naime, jednostavno rečeno, strukturne su pogreške one koje utječu na strukturu rečenice i samim time na provođenje sintaktičke analize. Nestrukturne pogreške, s druge strane, ne utječu na prvo od toga, a time i u manjoj mjeri ili nikako na drugo.

U strukturne pogreške ubrajamo pogreške u redu riječi i promjenu oblika riječi iz jedne u drugu gramatičku kategoriju. U nestrukturne pogreške pak ubrajamo svako neslaganje riječi koje se jedna na drugu odnose u rodu, broju, padežu i licu, odnosno pogreške u sročnosti.

4. Gramatički i stilski provjernik za hrvatski jezik

U okviru rada razrađeni su i iskušani različiti postupci namijenjeni pronalaženju gramatičkih i stilskih pogrešaka u tekstovima na hrvatskom jeziku. Konačni je rezultat sustav koji kombinira statistički pristup i pristup temeljen na pravilima. U ovome poglavlju opisani su svi uporabljeni postupci, bez obzira na njihovu uspješnost.

4.1. Jezični model

4.1.1. O jezičnim modelima općenito

Jezični modeli sadrže informaciju o tome koliko je vjerojatno da će se određeni slijed riječi (ili nekih drugih jezičnih entiteta) pojaviti u nekom tekstu. Tu informaciju crpe iz velikih korpusa jezično ispravnoga teksta – što se neki slijed riječi više puta pojavio u korpusu, to je njegova vjerojatnost veća i obratno. Broj pojavljivanja slijeda riječi u korpusu (odnosno njegovu vjerojatnost u jezičnom modelu) možemo dovesti u vezu s njegovom jezičnom ispravnosti. Ako je korpus dovoljno velik, a neki se slijed riječi u njemu uopće nije pojavio ili se pojavio veoma mali broj puta, postoji mogućnost da je taj slijed riječi jezično neispravan.

Slijedove riječi (ili drugih jezičnih entiteta) u kontekstu jezičnog modela nazivamo n-gramima, pri čemu n označava broj riječi u slijedu. Ako n-gram zapišemo kao $w_1 w_2 \dots w_n$, ili skraćeno w_1^n, onda njegovu vjerojatnost možemo zapisati kao

$$P(w_1 w_2 \dots w_n)$$

Iz toga prema pravilu ulančavanja vjerojatnosti slijedi:

$$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2)\cdots P(w_n|w_1^{n-1})$$

Za velike n-grame izračunavanje bi i pamćenje vrijednosti poput $P(w_n|w_1^{n-1})$ zahtijevalo ogroman korpus i zauzimalo bi previše memorije pa se stoga u praksi koriste manji n-grami. Dovoljno dobri rezultati često se mogu postići već i uz bigramski model, u kojemu se $P(w_n|w_1^{n-1})$ aproksimira sa $P(w_n|w_{n-1})$. Uz bigramski model vjerojatnost u gornjemu izrazu postaje

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$

Jezični model gradi se brojanjem n-grama u korpusu i normaliziranjem njihovih vjerojatnosti na raspon od 0 do 1. Normaliziranje se provodi tako da se ukupan broj pojavljivanja jednoga te istoga n-grama dijeli s brojem pojavljivanja svih n-grama koji s tim n-gramom dijele isti *prefiks*, odnosno slijed od prvih $n-1$ riječi. Kod bigramskog modela prefiks je samo jedna riječ, pa vjerojatnost bigrama definiramo na sljedeći način:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)}$$

. Taj se izraz može i pojednostavniti, s obzirom na to da je broj pojavljivanja bigrama koji počinju riječju w_{n-1} zapravo jednak broju pojavljivanja same riječi w_{n-1}:

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

4.1.2. Zaglađivanje

Osnovni problem koji se javlja kod jezičnog modela proizlazi iz neizbježne ograničenosti korpusa. Ma koliko velik bio korpus kojim raspolažemo, uvijek će postojati n-grami koji se u njemu nisu pojavili, a inače su sasvim prihvatljivi i mogu se pojaviti u nekom drugom tekstu. Takvi n-grami nazivaju se *nul n-grami* i njihova je vjerojatnost jednaka nuli – ako se nul n-gram pojavi u nekoj rečenici, čitava će rečenica zbog njega biti proglašena neispravnom, premda postoji mogućnost da je ispravna.

Problem ograničenosti korpusa djelomično se rješava metodom koja se naziva zaglađivanje (engl. *smoothing*). Zaglađivanjem se n-grami s vjerojatnošću nula ponovno vrednuju i po određenom im se principu dodjeljuje pozitivna vrijednost. Jedan jednostavan princip jest uvećavanje broja pojavljivanja za jedan svim n-gramima, i onima koji su se pojavili u korpusu i onima koji nisu, a mogli su s obzirom na sve različite riječi (obličnice) koje su se pojavile u korpusu. Međutim, na taj se način prevelika vjerojatnosna masa pridaje n-gramima koji se nisu

pojavili pa je preciznost modela znatno lošija. Nešto je složenija Witten-Bellova metoda zaglađivanja, no daje i bolje rezultate.

Witten-Bellovo zaglađivanje na n-grame s vjerojatnošću nula gleda kao na događaje koji se još nisu dogodili. Kada se pojave, bit će to prvi put da ih vidimo, pa se pri modeliranju njihove vjerojatnosti možemo poslužiti vjerojatnošću prvog opažanja bilo kojeg n-grama. Ta metoda i inače se često koristi u statističkoj obradi jezika, a vrlo je dobar način ublažavanja posljedica ograničenog korpusa. Naime, što je korpus manji, više će biti n-grama koje nismo nikad vidjeli, ali će ujedno i vjerojatnost novih događaja biti veća.

Kako bismo procijenili vjerojatnost da ćemo neki n-gram vidjeti prvi put, potrebno je prebrojiti koliko smo puta vidjeli novi tip n-grama u korpusu, što zapravo odgovara broju različitih tipova n-grama (budući da smo svaki od njih prvi put vidjeli točno jednom). Vjerojatnost novog događaja u skupu svih događaja jednaka je omjeru broja novih događaja i ukupnog broja svih događaja (događaji pronalaženja n-grama i događaji pronalaženja novog tipa n-grama). Ako s N označimo sve događaje pronalaženja n-grama, a s T događaje pronalaženja novog tipa n-grama, onda ukupna vjerojatnost svih n-grama s vjerojatnošću nula glasi:

$$\sum_{i:c_i=0} p_i^* = \frac{T}{N+T}$$

Ukupnu vjerojatnost danu gornjim izrazom potrebno je razdijeliti na pojedine n-grame. Jedna od mogućnosti jest da ukupnu vjerojatnost podijelimo na ravnomjerne dijelove s obzirom na broj n-grama s vjerojatnošću nula (označen sa Z):

$$p_i^* = \frac{T}{Z(N+T)}$$

Vjerojatnosna masa koju smo pridodali n-gramima s vjerojatnošću nula morala je od negdje doći, stoga vjerojatnost viđenih n-grama definiramo na sljedeći način:

$$p_i^* = \frac{c_i}{Z(N+T)} \; ako \; je \; (c_i > 0)$$

Još je jedan način na koji možemo razdijeliti vjerojatnost na nul n-grame – umjesto da na različite nul n-grame gledamo kao na ravnopravne događaje, u dodjeljivanju vjerojatnosti možemo u obzir uzeti informaciju o njihovu prefiksu. Tako ćemo nul n-gramima čiji je prefiks češći i pojavljuje se u većem broju tipova n-grama dodijeliti veći dio vjerojatnosne mase. U slučaju bigramskog modela, ukupna vjerojatnost svih nul bigrama sada se definira s obzirom na prefiksnu

riječ w_x ($T(w_x)$ je broj tipova bigrama koji počinju riječju w_x, a $N(w_x)$ je broj pojavljivanja svih bigrama koji počinju tom istom riječju):

$$\sum_{i:c(w_x w_i)=0} p^*(w_i|w_x) = \frac{T(w_x)}{N(w_x) + T(w_x)}$$

Vjerojatnosnu masu raspoređujemo jednoliko na sve moguće nul bigrame $w_{i-1}w_i$ s istim prefiksom w_{i-1}, a takvih ima $Z(w_{i-1})$ (svi bigrami sastavljeni od prefiksa w_{i-1} i riječi iz korpusa koji se u tom obliku nisu pojavili u korpusu):

$$p^*(w_i|w_{i-1}) = \frac{T(w_{i-1})}{Z(w_{i-1})(N(w_{i-1}) + T(w_{i-1}))}$$

Konačno, vjerojatnosti viđenih bigrama ublažavamo na isti način:

$$\sum_{i:c(w_x w_i)>0} p^*(w_i|w_x) = \frac{c(w_x w_i)}{c(w_x) + T(w_x)}$$

4.1.3. Uloga u provjeri

Jezični model u provjeri možemo primijeniti na dva načina: s oznakama vrste riječi i sa samim riječima. Budući da je riječ o statističkoj metodi, u objema je slučajevima potreban korpus nad kojim će se jezični model izgraditi, a u prvom je slučaju potreban i označivač vrsta riječi. Ideja je da se rečenice proglase ispravnima ili neispravnima u ovisnosti o tome je li umnožak vjerojatnosti njihovih n-grama veći od postavljenog praga. Kako bi se smanjio broj lažnih uzbuna, primjenjuje se Witten-Bellovo zaglađivanje (v. 4.1.2.). Jezični model s riječima razlikuje se od onog s oznakama vrste riječi u dosegu – trebao bi otkriti puno više grešaka od jezičnog modela s oznakama, ali će polučiti i puno više lažnih uzbuna. Osim za provjeru na razini rečenica, informaciju o vjerojatnostima rečeničnih n-grama možemo iskoristiti i pri otkrivanju grešaka na lokalnoj razini. Ako je moguće utvrditi točnu poziciju pogreške, tada s pomoću jezičnog modela s riječima možemo predložiti potencijalne ispravke.

4.2. Provjera n-grama

Provjera n-grama nalik je provjeri temeljenoj na jezičnom modelu, ali je jednostavnija. Cilj je pronaći sve n-grame u tekstu koji se niti jednom nisu pojavili korpusu i označiti ih kao potencijalno pogrešne. Na taj način pobliže se određuje pozicija pogreške, za razliku od provjere na razini rečenice kod jezičnog modela.

Implementirana su dva različita tipa n-grama, n-grami s oznakama vrste riječi i hibridni n-grami. Prvi se sastoje isključivo od oznaka vrste riječi, dok se u drugima u određenom redoslijedu mogu pojaviti i riječi i oznake njihove vrste.

4.2.1. N-grami s oznakama vrste riječi

N-grami koji se sastoje samo od oznaka vrste riječi korisna su apstrakcija kojom se možemo poslužiti pri provjeri. U označavanju je korišten označivač opisan u (Osmann, 2011), koji podržava ukupno 42 oznake. Time se u odnosu na n-grame sastavljene od riječi, čijih oblika u jeziku ima na stotine tisuća, znatno smanjuje broj kombinacija koje tvore n-gram. Primjerice, za kvadrigram je broj mogućih kombinacija 42^4, a broj kombinacija koje mogu tvoriti ispravnu sintagmu vjerojatno je puno manji. Tako je i velik broj takvih kombinacija koje potencijalno tvore ispravnu sintagmu moguće pronaći na puno manjem korpusu s obzirom na n-grame sastavljene od riječi. Ipak, ni u ovom slučaju ne možemo računati da ćemo pronaći sve takve kombinacije i tako u potpunosti odijeliti sigurno neispravne od potencijalno ispravnih kombinacija. S druge se strane javlja i problem koji kod n-grama sastavljenog od riječi ne postoji – tamo, naime, ne govorimo o „potencijalno ispravnim“ kombinacijama riječi, s obzirom na to su kombinacije riječi koje su jednom bile ispravne uvijek ispravne (pod uvjetom da je čitav korpus ispravan). Ovdje se pak može pojaviti slijed oznaka koji se istovremeno može primijeniti i na ispravne i na neispravne sintagme. Također, na provjeru se negativno može odraziti i činjenica da provjernik ne radi sa stopostotnom točnošću.

Primjer izvođenja postupka možemo vidjeti na sljedećoj rečenici:

`Izrazio je nadu da će se liječnicima i drugim medicinskim radnicima osigurati jednaki uvjeti.`

Rečenicu najprije predajemo označivaču i dobivamo oznake kao u tablici 4.1. Analizom rečenice s obzirom na kvadrigrame oznaka utvrđujemo kako su se svi kvadrigrami pojavili u korpusu. Ako sada u rečenici uklonimo riječ *radnicima* i ponovno rečenicu analiziramo na jednak način, otkrit ćemo dva kvadrigrama koji se nisu bili pojavili u korpusu – to su kvadrigrami oznaka za sintagme *i drugim medicinskim osigurati* te *drugim medicinskim osigurati jednaki* (v. tablicu 4.2). Sve n-grame koji se nisu pojavili u korpusu smatramo pogrešnima, a ako se više takvih n-grama preklapa, sve ih stapamo u jednu pogrešku. S obzirom na to da se navedena dva kvadrigrama preklapaju, pogrešnom ćemo označiti sintagmu od pet riječi *i drugim medicinskim osigurati jednaki*:

`Izrazio je nadu da će se liječnicima <ERROR> i drugim medicinskim osigurati<\ERRO`
`jednaki uvjeti.`

Valja primijetiti da kvadrigram oznaka „`D GI Z I`“ kojim je označena neispravna sintagma *medicinskim osigurati jednaki uvjeti* postoji u korpusu i da se stoga smatra ispravnim. To je primjer situacije u kojoj se isti slijed oznaka može odnositi i na neispravne i na ispravne sintagme.

Tablica 4.1: Ispravna rečenica s oznakama.

Značke	**Oznake vrste**
Izrazio	GDR
je	GPOM
nadu	I
da	V
će	GPOM
se	Z
liječnicima	I
i	V
drugim	BR
medicinskim	D
radnicima	I
osigurati	GI
jednaki	Z
uvjeti	I
.	PKD

Postavlja se pitanje kakve sve vrste pogrešaka možemo otkriti n-gramima oznaka. Oznake vrste riječi donose visoku razinu apstrakcije, no i manju ekspresivnost. To znači da će nam za prikupljanje n-grama biti dovoljan manji korpus, ali da ćemo ujedno moći otkriti i manji broj različitih vrsta pogrešaka. Provjerom temeljenom na n-gramima oznaka cilja se, od pogrešaka koje su u domeni gramatičkih i stilskih provjernika (v. 3.), prvenstveno na pogreške u redu riječi, odnosno strukturne pogreške. Tu spadaju i pravopisne pogreške kojima se narušava struktura rečenice, poput nepravilne uporabe zareza. Ipak, ni od strukturnih pogrešaka neće se moći otkriti sve:

– ne mogu se otkriti one pogreške u kojima se isti slijed oznaka može primijeniti i na ispravnu i na neispravnu sintagmu jer se svaki takav slijed

Tablica 4.2: N-grami oznaka u neispravnoj rečenici (nedostaje riječ *radnicima*).

Značke	N-grami oznaka	N-gram iz korpusa
Izrazio je nadu da	GDR GPOM I V	+
je nadu da će	GPOM I V GPOM	+
nadu da će se	I V GPOM Z	+
da će se liječnicima	V GPOM Z I	+
će se liječnicima i	GPOM Z I V	+
se liječnicima i drugim	Z I V BR	+
liječnicima i drugim medicinskim	I V BR D	+
i drugim medicinskim osigurati	**V BR D GI**	**-**
drugim medicinskim osigurati jednaki	**BR D GI Z**	**-**
medicinskim osigurati jednaki uvjeti	D GI Z I	+
osigurati jednaki uvjeti .	GI Z I PKD	+

oznaka smatra ispravnim;

– ne mogu se otkriti pogreške koje nisu vidljive lokalno, na razini n-grama određene veličine.

Osim toga, mogu se pojaviti i lažne uzbune, u slučaju kad se inače ispravni slijed oznaka nije pojavio u korpusu. Takve su situacije ipak rjeđe jer mogućih kombinacija oznaka ima neusporedivo manje nego kombinacija riječi.

Budući da se apstrahiranjem teksta na oznake vrste riječi gubi informacija o oblicima riječi, n-gramima oznaka ne mogu se otkriti morfološke pogreške, odnosno pogreške u sročnosti. Dakako, gubi se i informacija o značenju riječi, čime je isključeno i otkrivanje semantičkih pogrešaka. Preostaju još stilske pogreške različitih vrsta – premda ni one nisu u fokusu ovakve provjere, moguće je da se njome otkriju pojedine stilske pogreške vezane uz red riječi.

4.2.2. Hibridni n-grami

Hibridni n-grami mogu istovremeno biti sastavljeni i od riječi (tj. znački) i od oznaka vrste riječi. Time se smanjuje razina apstrakcije i povećava ekspresivnost, ali i potreba za većim korpusom. U programu je omogućeno sastavljanje n-grama na proizvoljan način, primjenom uzorka kojim se definira koliko će u n-gramu biti znački, a koliko oznaka, i gdje će se koje nalaziti. Što je u n-gramu više znački, a manje oznaka, to je provjera stroža i otkriva više pogrešaka, ali je i broj lažnih

uzbuna veći. Tako se zapravo hibridni n-grami po ekspresivnosti nalaze između n-grama sastavljenih isključivo od oznaka i onih sastavljenih isključivo od riječi.

Postupak provjere odvija se na gotovo jednak način kao i kod n-grama oznaka. Rečenice se najprije u cijelosti označe oznakama vrste riječi, a zatim se analiziraju s obzirom na definirani tip hibridnoga n-grama. Osnovna je razlika što se ne provjeravaju svi n-grami, nego samo oni kod kojih su se sve riječi (značke) u tom obliku pojavile u korpusu (pritom velika i mala slova nisu bitna). Razlog za to jest taj da se ublaži utjecaj ograničenosti korpusa i smanji broj lažnih uzbuna. U tablici 4.3 prikazana je analiza rečenice iz odjeljka 4.2.1. s obzirom na hibridni n-gram koji se sastoji redom od jedne oznake, jedne značke i zatim još jedne oznake (*oznaka-značka-oznaka*).

Tablica 4.3: Hibridni n-grami tipa *oznaka-značka-oznaka* u neispravnoj rečenici (nedostaje riječ *radnicima*).

Značke	**Hibridni n-grami**	**N-gram iz korpusa**
Izrazio je nadu	GDR je I	+
je nadu da	GPOM nadu V	+
nadu da će	I da GPOM	+
da će se	V će Z	+
će se liječnicima	GPOM se I	+
se liječnicima i	**Z liječnicima V**	**-**
liječnicima i drugim	I i BR	+
i drugim medicinskim	V drugim D	+
drugim medicinskim osigurati	**BR medicinskim GI**	**-**
medicinskim osigurati jednaki	**D osigurati Z**	**-**
osigurati jednaki uvjeti	GI jednaki I	+
jednaki uvjeti .	Z uvjeti PKD	+

Primjenom hibridnih n-grama u ponekim se slučajevima mogu izbjeći posljedice „pregrube“ podjele u označavanju,[1] zbog koje se javljaju sljedovi oznaka koji su istodobno primjenjivi i na ispravne i na neispravne sintagme. Osim toga, moguće je i ublažiti posljedice eventualnih problema u slučaju kad se označivač

[1]Pod *pregrubom podjelom* misli se na situaciju u kojoj su istom oznakom obuhvaćene riječi koje pripadaju istoj gramatičkoj kategoriji, ali različitim podkategorijama, odnosno situaciju kad riječi nisu do kraja razgraničene s obzirom na svoje različitosti. Npr., zamjenice možemo sve označiti istom oznakom, ali ih možemo i dodatno podijeliti po tipu na osobne, povratne, posvojne, itd.

pri radu oslanja na kontekst; npr., u sintagmi *u toj važno izjavi*, u kojoj je riječ *važnoj* pogrešno napisana kao *važno* (zbog čega je osjećamo više kao prilog nego kao pridjev), označivač bi mogao spornu riječ svejedno označiti kao pridjev i tako zamaskirati pogrešku. Ako je u hibridnom n-gramu sadržano više od jedne riječi, provjerom će biti obuhvaćene i pogreške u morfologiji odnosno sročnosti za uključene riječi. Prednost nad n-gramima koji se sastoje isključivo od riječi, s kojima je također moguće provjeravati sročnost, jest što je ovdje moguće to učiniti za nesusjedne riječi, a da između njih stoje samo oznake kao manje ograničavajući elementi (npr. kao u hibridnom n-gramu tipa *značka-oznaka-značka*).

4.3. Provjera sročnosti među imenicama i pridjevima

Imenice i pridjevi koji se na njih odnose moraju se prema pravilima sročnosti slagati u rodu, broju i padežu. Nije nimalo jednostavno odrediti na koju se imenicu (ili više njih) odnosi pridjev, ali čest je slučaj da se odnosi na imenicu koja slijedi iza njega neposredno ili na nekoliko riječi udaljenosti. Ta je činjenica iskorištena pri definiranju pravila kojim se provjerava sročnost među imenicama i pridjevima. Pritom je za dobivanje informacije o obliku imenica i pridjeva uporabljen lematizator, alat opisan u (Šnajder, 2010). Lematizator između ostaloga za određeni oblik riječi može dati sve moguće morfosintaktičke opise koji bi se na njega mogli odnositi. Tako se za svaki pridjev i imenicu koja slijedi nakon njega provjerava postoji li barem jedna dozvoljena kombinacija morfosintaktičkih opisa (ona u kojoj su rod, broj i padež isti). Ako se ne pronađe niti jedna takva kombinacija, na pridjev i imenicu upozorava se kao pogrešne. Takav je pristup razmjerno jednostavan i može polučiti velik broj lažnih uzbuna, ali je njime moguće otkriti brojne pogreške u sročnosti i upozoriti na njih.

5. Eksperimentalno vrednovanje

U ovome poglavlju opisani su korpusi i alati uporabljeni pri izgradnji postupaka provjere i njihovu vrednovanju, a nakon toga dani su i komentirani rezultati vrednovanja.

5.1. Korpusi i alati

Pri učenju jezičnog modela i prikupljanju različitih tipova n-grama iskorišten je korpus koji se sastoji od članaka iz Vjesnika, političkoga dnevnoga lista koji izlazi u Zagrebu od 1940. godine. Korpus sadržava sva Vjesnikova dnevna izdanja u razdoblju od 31. svibnja 1999. do 1. studenog 2009. godine, a sastoji se od sveukupno 4.466.178 rečenica. Sve se rečenice nalaze u jednoj datoteci i odvojene su u zasebne redove.

U većini postupaka zahtijevale su se i oznake vrste riječi iz korpusa, pa je jedan manji dio korpusa označen. Pri označavanju, a i kasnije pri ispitavanju, uporabljen je označivač vrsta riječi opisan u (Osmann, 2011). Njime je označeno prvih 17.000 rečenica iz korpusa, što se naknadno pokazalo nedostatnim za pojedine postupke. Razlog takog malog broja označenih rečenica jest taj što je označavanje trajalo predugo – za označavanje jedne rečenice potrebno je u prosjeku između dvije i tri sekunde, što znači da bi označavanje jednog milijuna rečenica, koliko je prvotno bilo planirano, trajalo između 25 i 35 dana. Budući da toliko dugačko razdoblje za označavanje nije bilo raspoloživo, u ispitavanju nije bilo druge nego koristiti se manjim brojem označenih rečenica.

U ispitivanju je uporabljen samostalno sastavljen korpus dan u dodatku A. Korpus je sastavljen od rečenica koje su se pojavile u nedavnim člancima objavljenim na internetskim stranicama Vjesnika. Rečenice su prilagođene i u njih su ubačene raznovrsne gramatičke i stilske pogreške, a u ispitivanju različitih postupaka iskorištene su samo one rečenice koje sadržavaju tip pogrešaka kojemu je postupak namijenjen (npr., u ispitivanju provjere s pomoću n-grama oznaka

nisu iskorištene rečenice koje sadrže pogreške u sročnosti; v. 4.2.1.). Razlog odabira Vjesnikovih članaka kao izvora za prikupljanje rečenica jest što su statističke metode, kakve su rabljene u ovom radu, osjetljive na promjenu domene teksta. Čak i male promjene u stilu i vokabularu, kakve se mogu pojaviti i u tekstovima pisanim s istom svrhom, mogu pogoršati rezultate statističkih metoda. Potrebno je stoga i učenje statističkih informacija, kao i ispitivanje nakon toga, obavljati na što specijaliziranijem tipu tekstova.

U osmišljavanju gramatičkih i stilskih pogrešaka nastojalo se što više približiti realnim pogreškama, onima koje se mogu dogoditi i koje ljudi rade svakodnevno. Pri tome se autor vodio svojim iskustvom u pisanju tekstova, ali i čitanju tuđih tekstova i uočavanju pogrešaka u njima. Pogreške su radi lakšeg snalaženja i razlikovanja označene metaatributima s obzirom na njihov tip. Metaatributima razlikujemo pogreške umetanja riječi ili sintagme, promjene riječi te umetanja ili brisanja razmaka (razdvajanje i stapanje riječi).

Osim Osmannova označivača, od dodatnih alata u radu je iskorišten i lematizator opisan u (Šnajder, 2010). Lematizator je sposoban raspoznati imenice, glagole i pridjeve i za njih može dati pune morfosintaktičke opise prema normi MULTEX-East.

5.2. Rezultati ispitivanja

U prikazu rezultata iskorištene su tri mjere uspješnosti – preciznost (engl. *precision*), odziv (engl. *recall*) i F-mjera. Da bismo definirali te mjere, potrebne su nam vrijednosti TP, FP i FN, koje uz TN sačinjavaju tablicu zabune (engl. *confusion matrix*) ili kontingencijsku tablicu (engl. *contingency table*). One u ovome kontekstu imaju sljedeće značenje:

TP (engl. *true positives*) – broj riječi ili sintagmi (općenito, dijelova u tekstu) koje su prepoznate kao pogrešne, a zaista i jesu pogrešne;

FP (engl. *false positives*) – broj dijelova u tekstu koji su prepoznati kao pogrešni, ali zapravo to nisu;

FN (engl. *false negatives*) – broj stvarnih pogrešaka koje nisu prepoznate niti označene.

Preciznost sada definiramo kao udio stvarno pogrešnih izraza u odnosu na sve izraze koji su kao takvi prepoznati:

$$Preciznost = \frac{TP}{TP + FP}$$

Odziv nam govori koliko je pogrešaka pronađeno od sveukupnog broja pogrešaka u korpusu za ispitivanje:

$$Odziv = \frac{TP}{TP + FN}$$

Kako su te mjere međusobno zavisne i ne možemo na njih gledati odvojeno (što više stremimo ka preciznosti, to ćemo više ugroziti odziv i obrnuto), definiramo F-mjeru kao njihovu harmonijsku sredinu:

$$F = \frac{2 \cdot Preciznost \cdot Odziv}{Preciznost + Odziv}$$

5.2.1. N-grami oznaka

Rezultati provjere s n-gramima oznaka prikazani su u dvama tablicama, 5.1 i 5.2. U provjeri je uporabljeno prvih 50 rečenica iz korpusa za vrednovanje, u kojima se pojavljuju isključivo strukturne pogreške. U tablici 5.1 u postotcima je dan udio upozoravajućih n-grama u ukupnom broju n-grama te udio korisnih n-grama u ukupnom broju upozoravajućih n-grama. Pod upozoravajućim n-gramima pritom se misli na one koji se nisu pojavili u korpusu za učenje, a pod korisnima na one od njih koji se zaista odnose na stvarnu pogrešku.

$$Upozoravajući = \frac{Broj\ n-grama\ koji\ se\ nisu\ pojavili\ u\ korpusu\ za\ učenje}{Ukupan\ broj\ n-grama\ u\ svim\ vrednovanim\ rečenicama} \cdot 100[\%]$$

$$Korisni = \frac{Broj\ n-grama\ koji\ se\ odnose\ na\ stvarnu\ pogrešku}{Ukupan\ broj\ upozoravajućih\ n-grama} \cdot 100[\%]$$

Te mjere bilo je nužno prikazati zbog načina na koji se odvija provjera – ako se pojave upozoravajući n-grami koji se međusobno preklapaju, na čitav se dio u tekstu počevši od prvog takvog n-grama pa sve do zadnjega gleda kao na jednu pogrešku. U slučaju kad ima puno upozoravajućih n-grama može se dogoditi da čitava rečenica ili njezin velik dio budu označeni kao pogrešni, što će neprirodno uvećati mjere preciznosti i odziva, dok će provjera istodobno postati neupotrebljiva. Zato nam samo mjere preciznosti i odziva nisu dovoljne, već uz njih moramo promotriti i postotke upozoravajućih i korisnih n-grama. Kako bismo provjeru zadržali u granicama upotrebljivost, dobro je udio upozoravajućih n-grama održati na maksimalno dvadesetak posto. Primjer potpuno neupotrebljive provjere, usprkos jako visokom odzivu i relativno visokoj preciznosti, jest ona s 6-gramima.

Tablica 5.1: Udjeli upozoravajućih n-grama u ukupnom broju n-grama te korisnih n-grama u ukupnom broju upozoravajućih n-grama za različite tipove n-grama oznaka.

Tip n-grama	Upozoravajući [%]	Korisni [%]
2-grami	0,19	50,00
3-grami	0,92	77,77
4-grami	6,12	61,40
5-grami	20,29	48,60
6-grami	45,31	40,32

U tablici 5.2 dane su mjere preciznosti i odziva te F-mjera. Iako se može činiti da je u provjeri najbolje koristiti 5-grame ili 6-grame, to s obzirom na udjele upozoravajućih i korisnih n-grama ipak nije tako. Štoviše, u tim se provjerama velik dio rečenica nepotrebno obuhvati u pogrešni dio, zbog čega gube na korisnosti. Najbolja kombinacija preciznosti, odziva i udjela upozoravajućih i korisnih n-grama jest ona kod 4-grama.

Tablica 5.2: Preciznost, odziv i F-mjera za različite tipove n-grama oznaka.

Tip n-grama	Preciznost [%]	Odziv [%]
2-grami	50,00	2,00
3-grami	85,71	13,64
4-grami	67,65	46,00
5-grami	55,56	70,00
6-grami	62,12	82,00

5.2.2. Hibridni n-grami

Provjera s hibridnim n-gramima vrednovana je na sličan način kao i ona s n-gramima oznaka. Kao dodatna informacija dan je i udio preskočenih n-grama, odnosno n-grama koji, zbog nepojavljivanja neke od njihovih znački u korpusu za učenje, nisu sagledavani u provjeri:

$$Preskočeni = \frac{Broj\ n-grama\ koji\ su\ preskočeni\ u\ provjeri}{Ukupan\ broj\ n-grama\ u\ svim\ vrednovanim\ rečenicama} \cdot 100[\%]$$

I kod hibridnih n-grama slična je situacija s odnosom udjela upozoravajućih i korisnih n-grama naspram preciznosti i odziva. Iako provjera s n-gramima tipa

Tablica 5.3: Udjeli preskočenih i upozoravajućih n-grama u ukupnom broju n-grama te korisnih n-grama u ukupnom broju upozoravajućih n-grama za različite tipove hibridnih n-grama.

Tip n-grama	**Preskočeni** [%]	**Upozoravajući** [%]	**Korisni** [%]
značka-oznaka	9,50	16,28	13,10
oznaka-značka	9,12	14,05	17,24
oznaka-značka-oznaka	9,57	32,48	19,44

oznaka-značka-oznaka daje visok odziv, ona je ipak s praktičnoga aspekta beskorisna. Nažalost, rezultati ispitivanja s hibridnim n-gramima ne mogu se smatrati u potpunosti vjerodostojnima zbog ograničena korpusa na kojemu su hibridni n-grami prikupljani (samo 17.000 rečenica). Kako bi se postigli realni (i vjerojatno puno bolji) rezultati, potrebno je označiti i iskoristiti veći dio korpusa.

Tablica 5.4: Preciznost, odziv i F-mjera za različite tipove hibridnih n-grama.

Tip n-grama	**Preciznost** [%]	**Odziv** [%]
značka-oznaka	13,73	41,18
oznaka-značka	18,33	43,14
oznaka-značka-oznaka	24,54	80,00

6. Zaključak

U okviru ovoga diplomskog rada razrađeni su i iskušani različiti postupci namijenjeni provjeri gramatičke i stilske ispravnosti u tekstovima na hrvatskome jeziku. Postupci se temelje na dvama pristupima, na statistici i na pravilima. Statistički postupci koji su implementirani uključuju jezični model, n-grame oznake i hibridne n-grame, dok se na pravilima temelji provjera sročnosti među imenicama i pridjevima.

Jezični model izgrađen nad oznakama vrste riječi i s primjenjenim Witten-Bellovim zaglađivanjem nije dao očekivane rezultate – kako god se postavio prag između rečenica koje se smatraju ispravnima i onih koje se smatraju neispravnima, broj netočno označenih rečenica i dalje je bio visok. Najbolji rezultati postignuti su s n-gramima oznaka, prilično jednostavnim, ali i efikasnim pristupom. Na vrlo malom korpusu za učenje prikupio se dovoljan broj n-grama da se omogući relativno uspješno razgraničavanje između ispravnih i neispravnih konstrukcija. Najviše je podbacio pristup s hibridnim n-gramima, no tomu je u velikoj mjeri uzrok premali korpus.

U tom smjeru idu i prijedlozi za daljnja poboljšanja – trebalo bi označiti i iskoristiti veći dio korpusa, a u ubrzavanju tog postupka mogle bi se iskoristiti metode paralelnoga programiranja. S većim bi korpusom i rezultati postupaka temeljenih na hibridnim n-gramima i jezičnom modelu vjerojatno dali puno bolje rezultate. Osim toga, trebalo bi poraditi na kombiniranju ovih postupaka u jedan sustav te eventualnoj ciljanoj primjeni pravila za umanjivanje broja lažnih uzbuna u statičkim postupcima.

Literatura

Md. Jahanagir Alam, Naushad UzZaman, i Mumit Khan. N-gram Based Statistical Grammar Checker for Bangla and English. *Proceedings of Ninth International Conference on Computer and Information Technology (ICCIT 2006)*, 2006.

Antti Arppe. Developing a Grammar Checker for Swedish. U *The 12th Nordic Conference of Computational Linguistics*, stranice 13 – 27. Citeseer, 2000.

Emily M. Bender, Dan Flickinger, Stephan Oepen, Annemarie Walsh, i Timothy Baldwin. Arboretum: Using a Precision Grammar for Grammar Checking in CALL. U *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, stranice 83 – 86. Citeseer, 2004.

Andrew Bredenkamp, Berthold Crysmann, i Mirela Petrea. Looking for Errors: A Declarative Formalism for Resource-Adaptive Language Checking. U *Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens, Greece*. Citeseer, 2000.

Flora Ramírez Bustamante i Fernando Sánchez León. GramCheck: A Grammar and Style Checker. U *Proceedings of the 16th Conference on Computational Linguistics (COLING '96) – Volume 1*, stranice 175 – 181. Association for Computational Linguistics, 1996.

Johan Carlberger, Rickard Domeij, Viggo Kann, i Ola Knutsson. A Swedish Grammar Checker, 2002.

Lionel Clément, Kim Gerdes, i Renaud Marlet. A grammar correction algorithm: Deep parsing and minimal corrections for a grammar checker. U *Proceedings of the 14th International Conference on Formal Grammar (FG '09)*, stranice 47 – 63. Springer, 2011.

Berthold Crysmann, Nuria Bertomeu, Peter Adolphs, Dan Flickinger, i Tina Klüwer. Hybrid Processing for Grammar and Style Checking. U *Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1*, stranice 153 – 160. Association for Computational Linguistics, 2008.

Jens Eeg-Olofsson i Ola Knutsson. Automatic Grammar Checking for Second Language Learners – the Use of Prepositions. U *Proceedings of NoDaLiDa*. Citeseer, 2003.

Gerhard Fliedner. A System for Checking NP Agreement in German Texts. U *Proceedings of the ACL Student Research Workshop*, stranice 12 – 17, 2002.

Mandeep Singh Gill i Gurpreet Singh Lehal. A Grammar Checking System for Punjabi. U *22nd International Conference on Computational Linguistics*, COLING '08, stranice 149 – 152, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=1599288.1599290`.

Sylvana Sofkova Hashemi. Detecting Grammar Errors in Children's Writing: A Finite State Approach. U *Proceedings of the 13th Nordic Conference on Computational Linguistics (NoDaLiDa '01), Uppsala, Sweden*, 2000.

Anna S. Hein. A Chart-Based Framework for Grammar Checking (Initial Studies). U *11th Nordic Conference in Computational Linguistic (NoDaLiDa '98)*, stranice 68 – 80, 1998.

Verena Henrich i Timo Reuter. LISGrammarChecker: Language Independent Statistical Grammar Checking. Magistarski rad, Hochschule Darmstadt, Reykjavík University, 2009.

Tomáš Holan, Vladislav Kuboň, i Martin Plátek. A Prototype of a Grammar Checker for Czech. U *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*, stranice 147 – 154. Association for Computational Linguistics, 1997.

Philip S. Kernick i David M. Powers. A Statistical Grammar Checker, 1996.

Jorge Kinoshita, Laís do Nascimento Salvador, i Carlos E. D. de Menezes. CoGrOO: A Brazilian-Portuguese Grammar Checker Based on the CETENFOLHA Corpus. U *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC*, 2006.

Daniel Naber. A Rule-Based Style and Grammar Checker. Magistarski rad, Technische Fakultät, Universität Bielefeld, 2003. URL `http://www.languagetool.org/`.

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

Vjekoslav Osmann. Označavanje vrste riječi u tekstovima na hrvatskome jeziku. Magistarski rad, Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu, 2011.

Jong C. Park, Martha Palmer, i Gay Washburn. An English Grammar Checker as a Writing Aid for Students of English as a Second Language. U *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. Association for Computational Linguistics, 1997.

Antje Schmidt-Wigger. Grammar and Style Checking for German. U *Proceedings of the Second International Workshop on Control Language Applications (CLAW-1998)*, stranice 76 – 86. Citeseer, 1998.

Khaled F. Shaalan. Arabic GramCheck: A Grammar Checker for Arabic. *Software – Practice and Experience*, 35:643 – 665, June 2005. ISSN 0038-0644. doi: 10.1002/spe.v35:7. URL `http://portal.acm.org/citation.cfm?id=1067114.1067116`.

Stjepko Težak i Stjepan Babić. *Gramatika hrvatskoga jezika*. Školska knjiga, 11. izdanju, 1996.

Gregor Thurmair. Parsing for Grammar and Style Checking. U *Proceedings of the 13th Conference on Computational Linguistics (COLING '90) – Volume 2*, stranice 365 – 370. Association for Computational Linguistics, 1990.

Chae Young-Soog. Improvement of Korean Proofreading System Using Corpus and Collocation Rules. *Language, Information and Computation (PACLIC12)*, 18:328 – 333, 1998.

Dodatak A
Korpus za vrednovanje

U ozračju ulaska u EU izrazio je nadu da će se liječnicima i drugim medicinskim radnicima osigurati jednaki uvjeti za rad i znanost.
* U ozračju ulaska u EU izrazio je nadu da će se liječnicima i drugim medicinskim <MISSING=radnicima> osigurati jednaki uvjeti za rad i znanost.
U pozdravnom govoru podsjetio je da je Akademija jedna od najstarijih i najuglednijih medicinskih akademija u svijetu, krovna institucija vrhunskih hrvatskih liječnika znanstvenika.
* U pozdravnom govoru podsjetio je da je Akademija jedna od najstarijih i najuglednijih medicinskih akademija u <MISSING=svijetu>, krovna institucija vrhunskih hrvatskih liječnika znanstvenika.
I većina je medija to javno poduprla, premda vodećim strukturama u mnogima od njih sigurno nije bilo nimalo ugodno.
* I većina je medija to javno poduprla, premda vodećim strukturama u <MISSING=mnogima> od njih sigurno nije bilo nimalo ugodno.
Ništa se nakon niza skandala i unatoč istrazi nije do kraja rasvijetlilo.
* Ništa se <MISSING=nakon> niza skandala i unatoč istrazi nije do kraja rasvijetlilo.
Iako je ove godine prvi put nakon pola stoljeća otvorena mogućnost fleksibilnog početka školske godine, što su prije tri mjeseca slavodobitno objavila dva ministra, nijedna županija nije iskoristila pruženu mogućnost.
* Iako je ove godine prvi put nakon pola stoljeća otvorena mogućnost fleksibilnog početka školske godine, što su prije tri mjeseca slavodobitno objavila dva <MISSING=ministra>, nijedna županija nije iskoristila pruženu mogućnost.
Kako doznajemo, poticaj za pomak nastavne godine trebao je doći od osnivača, a ne od škola.

* Kako doznajemo <ADDED>vidimo<\ADDED>, poticaj za pomak nastavne godine trebao je doći od osnivača, a ne od škola.
Nakon dugih ljetnih praznika školarci će se po novom kalendaru ponovo odmarati tijekom proljetnih praznika, skraćenih u odnosu na lanjske.
* Nakon <ADDED>poslije<\ADDED> dugih ljetnih praznika školarci će se po novom kalendaru ponovo odmarati tijekom proljetnih praznika, skraćenih u odnosu na lanjske.
U Ministarstvu su shvatili da od produljenja praznika neće biti ništa još kada su pripremali to rješenje.
* U Ministarstvu su shvatili <ADDED>su<\ADDED> da od produljenja praznika neće biti ništa još kada su pripremali to rješenje.
Kao odvjetnik, nisam mogao odbiti Sanaderov prijedlog da se uključim u njegovu obranu, a njegovu predmetu pristupam kao i svakom drugom.
* Kao odvjetnik, nisam mogao odbiti Sanaderov prijedlog da se uključim u njegovu obranu, a njegovu predmetu pristupam kao <ADDED>kao<\ADDED> i svakom drugom.
Oduvijek sam smatrala da mudrost dolazi ili sa starošću ili s čitanjem i s razumijevanjem ljudskih bića.
* Oduvijek sam smatrala da mudrost dolazi ili sa starošću ili <ADDED>kroz<\ADDED> s čitanjem i s razumijevanjem ljudskih bića.
To zahtijeva jačanje međunarodne konkurentnosti gospodarstva kroz unutarnje strukturne reforme.
* To <CHANGED>zahtjeva<\CHANGED> jačanje međunarodne konkurentnosti gospodarstva kroz unutarnje strukturne reforme.
Assange traži blokadu švedskog zahtjeva za izručenjem.
* Assange traži blokadu švedskog <CHANGED>zahtijeva<\CHANGED> za izručenjem.
Drugi su išli u smjeru ponavljanja originala, vjerno slijedeći modu međuratnog razdoblja.
* Drugi su išli u smjeru ponavljanja originala, vjerno <CHANGED>sljedeći<\CHANGED> modu međuratnog razdoblja.
Nakon pada jemenskog predsjednika Abdulaha Saleha svi se pitaju tko je sljedeći.
* Nakon pada jemenskog predsjednika Abdulaha Saleha svi se pitaju tko je <CHANGED>slijedeći<\CHANGED>.
Zvijezde su mu svijetleći na nebu pokazivale put.
* Zvijezde su mu <CHANGED>svjetleći<\CHANGED> na nebu pokazivale put.

Europska komisija daje zeleno svjetlo za pristup Hrvatske.
* Europska komisija daje zeleno <CHANGED>svijetlo<\CHANGED> za pristup Hrvatske.
U dijelu hrvatske javnosti osjeća se razočaranje što nismo jasnije i glasnije artikulirali tadašnje raspoloženje.
* U dijelu hrvatske javnosti osjeća se <SPACE_INS>razočaran je<\SPACE_INS> što nismo jasnije i glasnije artikulirali tadašnje raspoloženje.
Izrazio je uvjerenost da se dobrim pristupom i informiranjem građana o EU može takav rezultat i ostvariti.
* Izrazio je <CHANGED>uvjeren je<\CHANGED> da se dobrim pristupom i informiranjem građana o EU može takav rezultat i ostvariti.
Nakon toga benediktinke su hrvatskog predsjednika odvele u obilazak jednog od najstarijih samostana u Šibeniku, koji se već godinama obnavlja sredstvima iz benediktinskih europskih fondova, ali i donacijama.
* Nakon toga benediktinke su hrvatskog predsjednika odvele u obilazak jednog od najstarijih samostana u Šibeniku, <MISSING=koji> se već godinama obnavlja sredstvima iz benediktinskih europskih fondova, ali i donacijama.
Budući da je ukupno dodijeljeno oko 100 odličja, jasno je o kakvom je hrvatskom uspjehu riječ, ističe predsjednica Hrvatskog saveza inovatora.
* Budući <MISSING=da> je ukupno dodijeljeno oko 100 odličja, jasno je o kakvom je hrvatskom uspjehu riječ, ističe predsjednica Hrvatskog saveza inovatora.
U pozivnom centru na telefone su se javljale poznate osobe iz javnog, kulturnog i političkog života Hrvatske.
* U pozivnom centru na telefone su se javljale <ADDED>su se<\ADDED> poznate osobe iz javnog, kulturnog i političkog života Hrvatske.
Na temelju zakona o pravu na informacije, Gong je prije nekih tjedan dana tražio od Vlade da objavi kompletan sadržaj svih privremeno zatvorenih pregovaračkih poglavlja.
* Na temelju zakona o pravu na informacije, Gong je <SPACE_INS>pri je<\SPACE_INS> nekih tjedan dana tražio od Vlade da objavi kompletan sadržaj svih privremeno zatvorenih pregovaračkih poglavlja.
Sumnja se da su se tijekom javnog skupa ponašali na osobito drzak i neprimjeren način te da su svojim postupanjem doveli sudionike skupa u ponižavajući položaj.
* Sumnja se da su se tijekom javnog skupa ponašali na osobito drzak

i neprimjeren način <MISSING=te> da su svojim postupanjem doveli sudionike skupa u ponižavajući položaj.
Najavio je da će policija preventivno prije same povorke detaljno pregledati trasu kojom će se ona kretati te da će se svaku sumnjivu osobu provjeriti.
* Najavio je da će policija <ADDED>će<\ADDED> preventivno prije same povorke detaljno pregledati trasu kojom će se ona kretati te da će se svaku sumnjivu osobu provjeriti.
Prije prošlih parlamentarnih izbora hrvatska se politička elita pomamila za blogovima, ali su u međuvremenu svi ugašeni ili barem toliko neaktivni da od njih nema koristi.
* Prije prošlih parlamentarnih izbora hrvatska se politička elita pomamila za blogovima, ali su u međuvremenu svi ugašeni ili barem toliko <SPACE_INS>ne aktivni<\SPACE_INS> da od njih nema koristi.
Otvara li nezadovoljstvo građana radom svih stranaka mogućnost za veći iskorak nekih političkih pokreta i pojedinaca?
* Otvara li nezadovoljstvo građana <CHANGED>radom<\CHANGED> svih stranaka mogućnost za veći iskorak nekih političkih pokreta i pojedinaca?
Ako netko ne napravi krivi potez, prosvjednici će se ispuhati, smatra dio politologa, dok drugi zaustavljanje prosvjeda očekuju s objavom datuma izbora.
* Ako netko ne napravi krivi potez, prosvjednici će se ispuhati, <ADDED>vjeruje<\ smatra dio politologa, dok drugi zaustavljanje prosvjeda očekuju s objavom datuma izbora.
Tamni oblaci nadvili su se nad američkim tržištem kapitala, a nema naznake da će ih išta ubrzo rastjerati.
* Tamni oblaci nadvili su se nad američkim <ADDED>nad<\ADDED> tržištem kapitala, a nema naznake da će ih išta ubrzo rastjerati.
Pale su i cijene dionica u tehnološkom sektoru, dok su u defenzivnim sektorima, kao što su uslužni i zdravstveni, porasle.
* Pale su i cijene dionica u tehnološkom sektoru, dok su u defenzivnim sektorima, kao što su uslužni i zdravstveni <MISSING=,> porasle.
Svjetska proizvodnja žitarica ove bi godine mogla narasti do rekordne razine zbog povećane sjetve i boljih prinosa, ali bi cijene trebale ostati visoke zbog niskih zaliha.
* Svjetska proizvodnja žitarica ove bi godine mogla narasti do rekordne razine zbog povećane sjetve i boljih prinosa, ali bi cijene <ADDED>bi<\ADDED>

trebale ostati visoke zbog niskih zaliha.
Nagli skok cijena nafte i nedavni brzi pad svjetskih zaliha žitarica povećavaju mogućnost nove krize u opskrbi hranom.
* Nagli skok cijena nafte i nedavni brzi pad svjetskih zaliha žitarica povećavaju mogućnost nove krize u <ADDED>u<\ADDED> opskrbi hranom.
S druge strane, u Organizaciji za ekonomsku suradnju i razvoj ističu da su danas globalne svjetske zalihe znatno veće nego prije dvije, tri godine.
* S druge strane, u Organizaciji za ekonomsku suradnju i razvoj ističu da <CHANGED>si<\CHANGED> danas globalne svjetske zalihe znatno veće nego prije dvije, tri godine.
Drugim riječima, zalihe pšenice lani su bile dovoljne do iduće žetve, ali postoji problem ako se u priču ubace svjetski mešetari koji mogu otkupiti velike količine pšenice i onda diktirati rast cijena.
* Drugim riječima, zalihe pšenice lani su bile dovoljne do <SPACE_INS>idu će<\SPACE_INS> žetve, ali postoji problem ako se u priču ubace svjetski mešetari koji mogu otkupiti velike količine pšenice i onda diktirati rast cijena.
Analitičari tvrde da će buduće cijene žitarica na globalnoj razini uvelike ovisiti i o povećanim potrebama najmnogoljudnijih zemalja poput Kine, Indije i Pakistana.
* Analitičari <CHANGED>tvrd<\CHANGED> da će buduće cijene žitarica na globalnoj razini uvelike ovisiti i o povećanim potrebama najmnogoljudnijih zemalja poput Kine, Indije i Pakistana.
Ne samo da će vam umjetnici pokazati kako rade, već će rado s vama i popričati.
* Ne samo da će vam umjetnici pokazati kako rade, već će <CHANGED>radi<\CHANGED> s vama i popričati.
Pod krošnjama kestenova pred vama nastaju pejzaži, portreti, mali umjetnički predmeti koje možete povoljno kupiti uz najljepši pogled na Zagreb.
* Pod krošnjama kestenova pred <ADDED>pred<\ADDED> vama nastaju pejzaži, portreti, mali umjetnički predmeti koje možete povoljno kupiti uz najljepši pogled na Zagreb.
Cjelodnevni program, s kojim se započelo već krajem svibnja, posjetiteljima pruža pregršt zabave.
* Cjelodnevni program, s kojim se započelo već krajem svibnja, posjetiteljima

<CHANGED>pruća<\CHANGED> pregršt zabave.
Prozračne poput paučine, hrvatske čipke ovih anonimnih dama nisu nimalo manje vrijedne od kakva kapitalnog slikarskog djela.
* Prozračne poput paučine, hrvatske čipke ovih anonimnih dama nisu nimalo <CHANGED>mane<\CHANGED> vrijedne od kakva kapitalnog slikarskog djela.
Doista, atmosfera je od samog početka bila vrlo prijateljska, orkestar je pri ulasku na podij bio pozdravljen pljeskom dobrodošlice, a odabrani je program mogao zainteresirati slušatelje s različitim glazbenim ukusima.
* Doista, atmosfera je od samog početka bila vrlo prijateljska, orkestar je pri ulasku na podij bio pozdravljen pljeskom dobrodošlice, a odabrani je program <ADDED>je<\ADDED> mogao zainteresirati slušatelje s različitim glazbenim ukusima.
Treba pričekati i vidjeti kako će ovaj novi koncept zaživjeti među domaćom publikom, no sigurno je kako žanr horora i znanstvene fantastike ima veliku i čvrstu bazu poklonika, koji će zasigurno biti zainteresirani za ovaj festival.
* Treba pričekati i vidjeti kako će ovaj novi koncept zaživjeti među domaćom publikom, no sigurno je kako žanr horora i znanstvene fantastike ima veliku i čvrstu bazu poklonika, koji će zasigurno biti zainteresirani <MISSING=za> ovaj festival.
Za razliku od prvog plakata, na kojemu sam mlad čovjek drskoga pogleda i s malo kosice, danas sam star i ćelav.
* Za razliku od prvog plakata, na kojemu sam mlad čovjek drskoga pogleda i s malo kosice <MISSING=,> danas sam star i ćelav.
Što se promijenilo, a što je ostalo isto u odnosu na vrijeme kad ste tvrdili da ne želite pokazati ništa novo ni originalno?
* Što se promijenilo, a što je ostalo isto u odnosu na vrijeme kad ste tvrdili da<SPACE_DEL>ne želite pokazati ništa novo ni originalno?
Poznato je da dobri umjetnici preko noći mogu postati loši i obrnuto, no o tome se uglavnom govori u pola glasa.
* <CHANGED>Poznat<\CHANGED> je da dobri umjetnici preko noći mogu postati loši i obrnuto, no o tome se uglavnom govori u pola glasa.
Mi smo mjera stvari, a unatoč tako važnoj poziciji, stalno nas se ignorira.
* Mi smo mjera stvari, a unatoč tako <CHANGED>važno<\CHANGED> poziciji, stalno nas se ignorira.
Slijede teme bezazlenih dječjih igara, s nejakom djecom voštanih vjeđa,

zatim motiv poklada, s malim maskiranim prosjacima i okrunjenom djecom, nadstvarni i sanjarski kartaši u prnjama.
* <CHANGED>Slijed<\CHANGED> teme bezazlenih dječjih igara, s nejakom djecom voštanih vjeđa, zatim motiv poklada, s malim maskiranim prosjacima i okrunjenom djecom, nadstvarni i sanjarski kartaši u prnjama.
Naporni treninzi znaju trajati i do osam sati dnevno, naročito zimi.
* Naporni treninzi znaju trajati i do osam sati dnevno <MISSING=,> naročito zimi.
Sve je u redu ako nema ozljeda i zato sam zadovoljan.
* Sve je u<SPACE_DEL>redu ako nema ozljeda i zato sam zadovoljan.
Šteta je što je izgorio olimpijski centar na Bjelolasici, tamo smo imali odlične uvjete i mnogi su sportaši zbog toga zakinuti.
* Šteta je što <MISSING=je> izgorio olimpijski centar na Bjelolasici, tamo smo imali odlične uvjete i mnogi su sportaši zbog toga zakinuti.
Baka je djevojčicu, koja je tada imala dvije godine, odvela u bolnicu na pregled.
* Baka je djevojčicu, koja <ADDED>koja<\ADDED> je tada imala dvije godine, odvela u bolnicu na pregled.
Prijašnja su ispitivanja pokazala da je prozvani protein odgovoran i za stvaranje novih folikula vlasi, iako se prije smatralo da je njihov broj zadan rođenjem.
* Prijašnja su ispitivanja pokazala da je prozvani protein odgovoran i za stvaranje novih folikula vlasi, iako se prije smatralo da<SPACE_DEL>je njihov broj zadan rođenjem.

STROJNA PROVJERA GRAMATIKE I STILA U TEKSTOVIMA NA HRVATSKOME JEZIKU

Sažetak

U tekstovima formalna karaktera zahtijeva se jezična ispravnost, što podrazumijeva usklađenost s gramatičkom, pravopisnom i rječničkom normom standardnoga jezika. Osim toga, pri pisanju takvih tekstova potrebno je i poštivati odlika dobra stila, koje se s obzirom na svrhu teksta mogu razlikovati. U izbjegavanju jezičnih pogrešaka pomoći nam mogu pravopisni, gramatički i stilski provjernici, no od provjernika koji podržavaju hrvatski jezik dostupni su zasada samo pravopisni. U okviru ovoga rada dan je pregled postupaka za ispravljanje gramatičkih i stilskih pogrešaka i iznesena je tipologija jezičnih pogrešaka u pisanju. Razrađeni su i iskušani postupci temeljeni na statistici i na pravilima namijenjeni pronalaženju gramatičkih i stilskih pogrešaka u tekstovima na hrvatskome jeziku. Ti su postupci opisani te su izneseni i komentirani rezultati njihova vrednovanja.

Ključne riječi: gramatički i stilski provjernik, pristup temeljen na statistici i pravilima, n-grami, jezični model

GRAMMAR AND STYLE CHECKER FOR CROATIAN LANGUAGE

Abstract

In various occasions it is required to produce texts that are written in formal style and that are correct regarding their spelling and grammar. There are several types of computer programs that can help us avoid language errors—namely spelling, grammar and style checkers—but there are no grammar and style checkers available for Croatian language. This paper gives an overview of techniques for correcting grammar and style mistakes and presents a typology of language errors in Croatian texts. Various statistical and rule-based techniques have been developed and described. Those techniques have also been evaluated and the results of the evaluation are presented and discussed.

Keywords: grammar & style checker, statistical and rule-based approach, n-grams, language model