Poređenje postupaka automatske morfološke anotacije tekstova na srpskom jeziku

Milan S. Sečujski, *Član, IEEE,* i Aleksandar D. Kupusinac

***Sadržaj* — U okviru rada ispitana je mogućnost ostvarivanja morfološke anotacije visoke tačnosti za srpski jezik. Istraživanje je obuhvatilo i izradu morfološkog rečnika srpskog jezika, kao i odgovarajućeg morfološki anotiranog korpusa tekstova na srpskom jeziku, na kome su eksperimenti vršeni. Implementirana su i upoređena tri pristupa: skriveni Markovljevi modeli, tehnike mašinskog učenja zasnovane na transformacionim pravilima, te pristup korišćenjem ekspertskog sistema, pri čemu je ispitan i uticaj određenih modifikacija postojećih metoda u cilju povećanja tačnosti. Najbolji rezultati postignuti su korišćenjem ekspertskog sistema, pri čemu su svi postignuti rezultati uporedivi sa rezultatima postignutim za druge jezike sa sličnim stepenom složenosti sistema morfoloških kategorija.**

***Ključne reči* — morfološka anotacija, obrada prirodnog jezika, korpusna lingvistika.**

I. Uvod

ZADATAK automatske morfološke anotacije jeste da za svaku reč u datom tekstu odredi kojoj vrsti reči ova pripada, kao i koje su vrednosti odgovarajućih morfoloških kategorija. Ovaj problem se javlja u gotovo svim aplikacijama jezičkih tehnologija, i za njegovo rešavanje do sada su korišćene razne metode, kao što su skriveni Markovljevi modeli i tehnike mašinskog učenja, pri čemu je znatno veća tačnost postignuta za jezike sa relativno siromašnim sistemom morfoloških kategorija. Srpski jezik, međutim, spada u grupu jezika sa vrlo složenom morfologijom, tako da su potrebne određene modifikacije postojećih metoda da bi se postigla tačnost približnija onoj postignutoj za neke druge jezike. Zadatak istraživanja opisanih u ovom radu jeste da se sistematski ispita mogućnost ostvarivanja morfološke anotacije visoke tačnosti na srpskom jeziku.

U drugom poglavlju objašnjen je pojam obrade prirodnog jezika i navedeni osnovni problemi koji se prilikom obrade prirodnog jezika javljaju. Uvedeni su pojmovi morfološke anotacije i morfološkog deskriptora, i ukazano je na osnovne pravce rešavanja problema morfološke anotacije. Pored toga, naznačeni su i osnovni problemi koji se tiču postupka morfološke anotacije kod jezika sa složenom morfologijom.

Istraživanja opisana u ovom radu delimično su finansirana od strane Ministarstva za nauku i tehnološki razvoj Republike Srbije u okviru projekta "Govorna komunikacija čovek-mašina" (TR-11001).

M. S. Sečujski, Fakultet tehničkih nauka, Novi Sad, Srbija; (telefon: 381-21-4852533; faks: 381-21-4752997; e-mail: secujski@uns.ac.rs).

A. D. Kupusinac, Fakultet tehničkih nauka, Novi Sad, Srbija; (telefon: 381-21-4852441; faks: 381-21-4752997; e-mail: sasak@uns.ac.rs).

Treće poglavlje sadrži opis jezičkih resursa neophodnih za istraživanja na polju automatske morfološke anotacije opisana u ovom radu. Ovi resursi obuhvataju modul za morfološku analizu koji se oslanja na obiman akcenatsko-morfološki rečnik srpskog jezika (oko 100.000 leksema, odnosno, oko 3,9 miliona pojedinačnih oblika reči), kao i morfološki anotirani korpus srpskog jezika, koji obuhvata tekstove različitog tipa i sadrži ukupno oko 200.000 reči.

U četvrtom poglavlju objašnjeno je na koji način se Markovljevi modeli mogu iskoristiti za automatsku morfološku anotaciju teksta i prikazani su rezultati eksperimenata u okviru kojih je morfološka anotacija teksta na srpskom jeziku izvedena na taj način.

U petom poglavlju objašnjeno je kako se algoritmi mašinskog učenja zasnovani na transformacionim pravilima mogu primeniti za automatsku morfološku anotaciju. Prikazani su rezultati eksperimenata u okviru kojih je tekst na srpskom jeziku anotiran na taj način. Osnovni algoritam je u izvesnoj meri prilagođen osobinama inflektivnih jezika, i izvršeno je poređenje performansi tako prilagođenog algoritma sa performansama osnovnog u zavisnosti od veličine korpusa za obuku.

U šestom poglavlju opisan je ekspertski sistem zasnovan na direktnoj implementaciji gramatičkih pravila srpskog jezika, koji se koristi i kao osnova za akcentuaciju teksta u okviru AlfaNum sintetizatora govora na osnovu teksta [1]. Prikazani su rezultati eksperimenta u okviru kog je automatska morfološka anotacija teksta na srpskom jeziku izvršena korišćenjem ovog sistema.

Tekst je zaključen sedmim poglavljem, u kom je izvršeno poređenje rezultata dobijenih korišćenjem različitih tehnika automatske morfološke anotacije, i gde je konstatovano da su u postojećim uslovima (u pogledu raspoloživih tehnika anotacije i raspoloživih jezičkih resursa) ekspertski sistemi i dalje superiorni u odnosu na sisteme koji morfološku anotaciju vrše u potpunosti automatski. Ovaj zaključak je u skladu i sa rezultatima dobijenim za određen broj drugih jezika.

II. Obrada prirodnog jezika i automatska morfološka anotacija

Obrada prirodnog jezika (eng. *natural language processing* – NLP) predstavlja oblast računarskih nauka koja se bavi interakcijom između ljudi i računara putem prirodnog[1] jezika. Obrada prirodnog jezika ispituje strategije pomoću kojih bi računarima bilo omogućeno da razumeju

[1] Termin „prirodni" u ovom kontekstu ukazuje na to da se radi o jeziku koji koriste ljudi u međusobnoj komunikaciji, a ne o formalnom ili računarskom jeziku.

i obrađuju prirodni jezik u njegovoj pismenoj ili usmenoj formi. *Razumevanje prirodnog jezika* predstavlja konverziju jezičkih celina, odnosno nizova reči prirodnog jezika, u odgovarajuću formalnu reprezentaciju informacije, koja je pogodnija za obradu od strane računara. Ova obrada može imati različite ciljeve, u zavisnosti od potreba konkretnog sistema. Sa druge strane, *generisanje prirodnog jezika* predstavlja pretvaranje na isti način formalno reprezentovane informacije u razumljivu jezičku celinu, odnosno niz reči. Uz termin *obrada prirodnog jezika*, često se, u istom ili sličnom značenju, može susresti i termin *računarska lingvistika* (eng. *computational linguistics*).

Statistička obrada prirodnog jezika oslanja se na stohastičke, probabilističke i statističke metode da bi rešila neke od navedenih problema, posebno kada su u pitanju duge rečenice kojima bi, na osnovu klasične gramatike, odgovaralo na stotine miliona mogućih stabala raščlanjivanja. Metode određivanja ispravne interpretacije oslanjaju se po pravilu na izuzetno obimne korpuse teksta i na metode kao što su, primera radi, Markovljevi modeli (eng. *Markov models* – MM). Statistička obrada prirodnog jezika obuhvata sve kvantitativne pristupe automatskoj obradi prirodnog jezika, uključujući probabilističko modelovanje, kao i znanja iz teorije informacija i linearne algebre [2]. Tehnologija statističke obrade prirodnog jezika oslanja se u velikoj meri na mašinsko učenje (eng. *machine learning*) i *data mining*, oblasti veštačke inteligencije koje obuhvataju učenje na osnovu obimnih skupova podataka. Ubrzanim razvojem računarske tehnologije, do koga je došlo poslednjih decenija, postignut je izuzetan napredak u ovoj oblasti.

Morfološka anotacija podrazumeva automatsko određivanje vrste reči za svaku reč u tekstu, odnosno, u širem smislu, određivanje pojedinih morfoloških svojstava svake reči u tekstu. Jasno je da se radi o problemu u izuzetnoj meri zavisnom od jezika, s obzirom na značajne razlike u morfologiji pojedinih jezika. U zavisnosti od složenosti morfologije, informacija koju je za svaku reč potrebno odrediti može biti različita. Primera radi, kod inflektivnih jezika, među koje se svrstava i srpski, morfološka anotacija može podrazumevati i određivanje vrednosti odgovarajućih morfoloških kategorija u slučaju promenljivih vrsta reči. Primera radi, u srpskom jeziku morfološke kategorije imenica su broj i padež, tako da je za svaku imenicu u tekstu potrebno odrediti i njih. Međutim, pored gramatičkih kategorija broja i padeža, imenice poseduju i gramatičku kategoriju roda, koja, premda nije morfološka već klasifikaciona, može biti od veoma velikog značaja za kasniju analizu teksta (primera radi, za sintaksnu analizu). Pored toga, često je od interesa znati i da li se radi o vlastitoj, zajedničkoj, zbirnoj, gradivnoj ili apstraktnoj imenici, tako da bi bilo korisno anotacijom obuhvatiti i tu informaciju. Jasno je, dakle, da morfološka anotacija u širem smislu može obuhvatiti i mnoge druge informacije pored vrste reči i vrednosti morfoloških kategorija, odnosno, da jedinstveni morfološki deskriptor ili *oznaka* (eng. *tag*), koji se pridružuje svakoj reči, može biti onoliko opširan i detaljan koliko to zahteva specifična primena anotiranog teksta. Kompleksnost morfoloških deskriptora (oznaka) može, prema tome, varirati u zavisnosti od namene sistema, ali je u principu manja kod jezika sa jednostavnijom morfologijom. Samim tim se kod takvih jezika može očekivati i veća tačnost algoritama za morfološku anotaciju.

Primera radi, za engleski jezik definisano je nekoliko standardnih skupova oznaka, kao što su Brownov skup oznaka, koji se sastoji od 179 jedinstvenih oznaka kojima je anotiran američki Brown korpus [3] i Penn Treebank skup oznaka, koji sadrži 45 oznaka i predstavlja pojednostavljenu verziju Brownovog skupa oznaka, a korišćen je za anotaciju Penn korpusa sintaksnih stabala [4]. Činjenica da broj oznaka u ovim skupovima nije velik posledica je relativne jednostavnosti engleske morfologije. Naime, imenice u engleskom jeziku mogu imati svega dva oblika – jedninu i množinu (*house/houses*), dok glagoli mogu imati pet (*write*/*writes*/*wrote*/*written*/*writing*). Odgovarajućih oblika u jezicima sa složenom morfologijom ima znatno više, tako da, primera radi, skup oznaka korišćen za anotaciju Praškog korpusa sintaksnih stabala [5], kao i Češkog nacionalnog korpusa [6] teorijski sadrži čak 3.030 oznaka, iako se svega trećina tog broja zaista javlja u čitavom korpusu. Zanimljivo je primetiti da se veoma učestalim funkcionalnim rečima pri konstrukciji sistema oznaka po pravilu dodeljuju sasvim posebne oznake, zbog njihovog specifičnog ponašanja u jeziku. Tako se, primera radi, u Brown korpusu glagol *do* ne obeležava standardnom oznakom za glagol u osnovnom obliku (VB), već ima posebnu oznaku (DO). Sličnim principom rukovode se i autori mnogih drugih sistema oznaka jer on doprinosi povećanju tačnosti automatske morfološke anotacije, kao i upotrebljivosti rezultata.

Prilikom anotacije inflektivnih jezika kao što su češki i srpski, po pravilu se koriste pozicioni sistemi oznaka, što znači da se morfološki deskriptori, odnosno oznake, zapravo sastoje od niza elementarnih oznaka, od kojih prva opisuje vrstu reči, dok se sledeće odnose na vrednosti pojedinih morfoloških kategorija i drugih kategorija od interesa.

A. Osnovni pristupi automatskoj morfološkoj anotaciji

Dve osnovne grupe savremenih sistema za automatsku morfološku anotaciju su (1) *ekspertski sistemi*, orijentisani na primenu pravila do kojih su došli stručni lingvisti, i (2) *automatski sistemi*, koji kroz samostalnu analizu obimnih tekstova pokušavaju da dođu do određenih zakonitosti koje bi kasnije primenili u anotaciji nepoznatog teksta [7].

Ekspertski sistemi su u potpunosti zavisni od jezika, tako da je primena sistema razvijenog za jedan jezik za morfološku anotaciju na drugom jeziku (čak i uz odgovarajuće modifikacije) moguća samo u slučaju veoma srodnih jezika. S druge strane, upravo sistemi iz ove grupe postižu najveću tačnost. Kao primer ekspertskog sistema koji postiže izuzetno visoku tačnost često se navodi EngCG (*English Constraint Grammar*), razvijen za engleski jezik [8], [9]. Ovaj sistem zasniva se na primeni konačnih automata realizovanih na osnovu ručno napisanih gramatičkih pravila, a u zavisnosti od vrste teksta koji se anotira, može imati tačnost i preko 99%.

S druge strane, u potpunosti automatski sistemi po

pravilu imaju nižu tačnost, ali za njihovu realizaciju nije potrebno uložiti veliki trud stručnih lingvista koji bi sastavljali gramatička pravila. Tekst koji se koristi za obuku sistema može biti unapred anotiran ili ne, u zavisnosti od čega se razlikuju nadgledane (eng. *supervised*) i nenadgledane (eng. *unsupervised*) tehnike. Nadgledane tehnike koriste unapred anotirani tekst kao osnovu za dobijanje informacija koje će se koristiti tokom anotacije, kao što su relativne učestalosti reči ili oznaka, verovatnoće pojave određenih sekvenca reči ili oznaka, kao i automatski identifikovana gramatička pravila. S druge strane, nenadgledane tehnike koriste znatno složenije matematičke algoritme kako bi otkrile sličnosti u ponašanju pojedinih reči i na osnovu toga ih grupisale po srodnosti, definišući time i sâm skup oznaka automatski, da bi zatim na osnovu dobijenih zakonitosti izvršile i klasifikaciju svake reči nepoznatog teksta. Nenadgledane tehnike su mnogo portabilnije od nadgledanih i njihova važna prednost je što se mogu koristiti i u odsustvu odgovarajućih jezičkih resursa. Međutim, pokazuje se da se veća tačnost anotacije generalno postiže korišćenjem nadgledanih tehnika, odnosno, da je u slučaju da anotirani korpus dovoljnog obima postoji bolje koristiti nadgledane tehnike (Merialdo, 1994).

III. Resursi za automatsku morfološku anotaciju na srpskom jeziku

Prvi korak u automatskoj morfološkoj anotaciji određene reči jeste određivanje spiska oznaka koje bi joj teoretski mogle biti pridružene. Od specifičnosti pojedinih jezika zavisi da li će se ovaj korak u manjoj ili u većoj meri oslanjati na odgovarajući morfološki rečnik. Primera radi, na jezicima sa relativno jednostavnom morfologijom (poput engleskog) rečnik ne mora postojati kao poseban lingvistički resurs, već je, u slučaju da je korpus anotiranog teksta koji će biti korišćen za obuku sistema za anotaciju dovoljno obiman, moguće sastaviti veoma obuhvatan rečnik jednostavnim evidentiranjem reči koje se pojavljuju u korpusu. Drugu krajnost predstavljaju jezici sa veoma složenom morfologijom (poput turskog), gde rečnik ne samo da mora postojati kao poseban lingvistički resurs, već je neophodno da pored toga budu implementirani određeni algoritmi za morfološku analizu, pošto se u tekstovima na ovakvim jezicima po pravilu javlja znatno veći procenat reči koje se ne nalaze u rečniku [10].

U okviru ovog istraživanja, korišćenjem za to posebno kreiranog softverskog alata, realizovan je AlfaNum akcenatsko-morfološki rečnik srpskog jezika, koji trenutno sadrži oko 100.000 leksema, odnosno, oko 3,9 miliona pojedinačnih oblika reči. Za istraživanje problema automatske morfološke anotacije tekstova na srpskom jeziku bilo je potrebno sastaviti i što obimniji korpus morfološki anotiranih rečenica. U okviru ovog istraživanja, korišćenjem za to posebno kreiranog softverskog alata, sastavljen je AlfaNum morfološki anotirani tekstualni korpus (ATC), koji sadrži blizu 11.000 rečenica sa ukupno oko 200.000 reči. Oznake u rečniku i korpusu međusobno su konzistentne, što znači da se eksperimenti nad korpusom mogu bez ikakvih ograničenja izvoditi uz oslanjanje na rečnik.

A. *Morfološki rečnik srpskog jezika*

AlfaNum morfološki rečnik realizovan je u okviru projekta razvoja govornih tehnologija, i njegova prvobitna namena bila je podrška automatskoj morfološkoj anotaciji u okviru sistema za sintezu govora na osnovu teksta na srpskom jeziku [1]. Iz tog razloga, pojedinačni unos u rečnik, pored morfološkog deskriptora u vidu niza oznaka morfoloških kategorija, sadrži i podatke o akcentuaciji svakog unosa, kao i podatke o osnovnom obliku reči, neophodne za lematizaciju teksta. Pod *unosom* se podrazumeva određeni oblik reči, zajedno sa podacima o osnovnom obliku te reči, vrednostima odgovarajućih morfoloških kategorija, kao i akcenatskoj konfiguraciji (nizu oznaka koji opisuje akcentuaciju svakog sloga). Primer unosa u AlfaNum rečnik bio bi:

Vb-p-1-- izgubićemo (izgubiti) [0\000].

Morfološke kategorije koje se obeležavaju zavisne su od vrste reči, tako da se, primera radi, kod glagola obeležavaju oblik, te rod, broj i lice, pri čemu se vrednosti roda, broja i lica obeležavaju samo ako su primenljive na dati oblik. U konkretnom primeru radi se o prvom licu (1) množine (p) glagola (V) *izgubiti* u futuru (b). S obzirom da futur nema različite oblike za pojedine rodove, kategorija roda u ovom primeru nije obeležena, kao ni kategorije padeža ni stepena poređenja, koje su aktuelne kod nekih drugih vrsta reči. Podaci o akcentuaciji reči dati su u vidu niza znakova koji opisuju položaj naglašenog sloga u reči i tip akcenta na njemu.

Poslednja verzija rečnika, na kojoj su izvršeni svi eksperimenti opisani u radu, sadrži ukupno 3.888.407 unosa, odnosno, 100.517 leksema. Unosima u rečnik pridruženo je ukupno 748 različitih oznaka. Statistika sadržaja rečnika prema vrstama reči predstavljena je u tabeli 1.

B. *Morfološki anotirani korpus srpskog jezika*

Kreiranje i obrada morfološki anotiranih tekstualnih korpusa predstavljaju skup i vremenski zahtevan proces, tako da je neprekidna potreba za njima problem sa kojim se susreće svaka jezička zajednica. Postojeći tekstualni korpusi na srpskom jeziku većinom nisu morfološki anotirani. Značajnije izuzetke predstavljaju *Korpus srpskog jezika* razvijen na Institutu za eksperimentalnu fonetiku i patologiju govora u Beogradu [11], koji obuhvata tekstove na srpskom jeziku iz perioda od 12. do 20. veka, i u kome je oko 11 miliona reči ručno anotirano, kao i srpski prevod romana „1984“. Džordža Orvela, koji predstavlja najznačajniji segment MULTEXT-East jezičkih resursa za srpski jezik [12] i obuhvata 108.805 reči. Međutim, zbog nekompatibilnosti oba navedena korpusa sa AlfaNum rečnikom, kao i potrebe za prisustvom tekstova različitog tipa u korpusu (uključujući i savremenu štampu), pokazalo se kao bolje rešenje kreirati novi korpus nego prilagođavati ijedan od postojećih korpusa rečniku (ili rečnik korpusu).

TABELA 1: SADRŽAJ ALFANUM MORFOLOŠKOG REČNIKA

vrsta reči	*broj leksema*	*broj unosa*
imenica (N)	46.408	460.816
zamenica (P)	77	2.974
glagol (V)	10.010	292.917
pridev (A)	36.807	3.108.355
broj (M)	281	8.684
prilog (R)	6.510	14.203
predlog (S)	105	131
veznik (C)	48	51
partikula (Q)	89	94
uzvik (I)	182	182
UKUPNO	100.517	3.888.407

TABELA 2: SADRŽAJ ALFANUM TEKSTUALNOG KORPUSA

vrsta reči	*učestalost*	*procentualni udeo*
imenica (N)	60142	30.07%
glagol (V)	37451	18.72%
pridev (A)	23903	11.95%
veznik (C)	20352	10.17%
predlog (S)	20064	10.03%
zamenica (P)	17347	8.67%
prilog (R)	9281	4.64%
broj (M)	7192	3.60%
partikula (Q)	3349	1.67%
rezidual (X)	848	0.42%
uzvik (I)	98	0.05%

AlfaNum morfološki anotirani korpus srpskog jezika obuhvata tekstove iz različitih izvora [13]. Najveći deo sadržaja (između 80% i 90%) obuhvata dnevna štampa, pri čemu su zastupljeni tekstovi iz različitih rubrika, uključujući politiku, ekonomiju, kulturu, sport, ali i rubrike kao što su zdravlje, slobodno vreme, zanimljivosti. Zastupljeni su i naučno-popularni tekstovi, putopisne reportaže i intervjui. Ostatak sadržaja korpusa čine prozni tekstovi različitih autora preuzeti sa Interneta, kao i određen broj tekstova enciklopedijskog karaktera. Osim u pogledu leksičkog sadržaja, vodilo se računa da korpus bude što reprezentativniji i u pogledu gramatičkog sadržaja, a posebno u pogledu zastupljenosti pojedinih glagolskih oblika.

Korpus u svom osnovnom formatu obuhvata, dakle, niz rečničkih unosa koji odgovara ispravnoj morfološkoj anotaciji rečenice. AlfaNum korpus anotiran je na nivou rečenice, što znači da se rečenice ne tretiraju kao deo jedinstvenog teksta, već se posmatraju svaka za sebe. Na ovaj način raskidaju se anaforske veze (veze sa prethodnim kontekstom), što povećava broj slučajeva u kojima je dopuštena mogućnost višestruke anotacije. Statistički opis sadržaja korpusa dat je u tabeli 2. Detaljniji podaci o sadržaju rečnika i korpusa, kao i o načinu definisanja oznaka, dati su u [14].

IV. AUTOMATSKA MORFOLOŠKA ANOTACIJA NA OSNOVU MARKOVLJEVIH MODELA

Skriveni Markovljevi modeli (eng. *Hidden Markov Models* – HMM) već dugo se vrlo uspešno koriste u različitim oblastima obrade prirodnog jezika, uključujući automatsko prepoznavanje govora i modelovanje sekvenci govornih činova. Skriveni Markovljev model je zapravo probabilistička funkcija Markovljevog procesa, odnosno lanca. Zanimljivo je da su Markovljevi procesi zapravo i razvijeni upravo u lingvističke svrhe – za modelovanje sekvenci slova u delima ruske književnosti [15], ali su već tada predstavljali potpuno univerzalan alat za statističku obradu podataka.

Osnovna pretpostavka na kojoj se Markovljevi modeli zasnivaju jeste da za predviđanje budućeg stanja nekog sistema nije potrebno ništa osim poznavanja njegovog sadašnjeg stanja.

Kod skrivenih Markovljevih modela nije poznata sekvenca stanja, već samo neka njena probabilistička funkcija. Model prelazi iz stanja u stanje, i u svakom stanju na osnovu određene probabilističke funkcije emituje *simbole* koji su vidljivi i čine *opservacionu sekvencu*, dok sama sekvenca stanja nije vidljiva. Pomenuta probabilistička funkcija definiše verovatnoće emitovanja pojedinih simbola u svakom stanju. Skriveni Markovljevi modeli su, dakle, korisni kada je potrebno modelovati određen nevidljiv niz događaja koji s određenim verovatnoćama generišu vidljiv niz događaja.

Dobar primer je upravo problem automatske morfološke anotacije. Niz tokena u tekstu koji treba anotirati je vidljiv, dok je niz oznaka koje odgovaraju tim tokenima nevidljiv. Zahvaljujući gramatičkim pravilima koja određuju strukturu rečenice, nisu svi nizovi oznaka podjednako verovatni. Takođe, pošto ni sve reči nisu podjednako učestale, neće se ni svaki token s podjednakom verovatnoćom emitovati u određenom stanju. Automatska morfološka anotacija na osnovu Markovljevih modela zasniva se na tome da se u prvom koraku dobije verna procena verovatnoćâ prelaza iz stanja u stanje (pri čemu jednom stanju može odgovarati i niz od više oznaka, što odgovara Markovljevim modelima višeg reda), kao i verovatnoćâ emitovanja pojedinih tokena u svakom stanju, a da se zatim na nepoznatom tekstu (nizu tokena) odredi ona sekvenca stanja za koju je najverovatnije da je generisala datu opservacionu sekvencu.

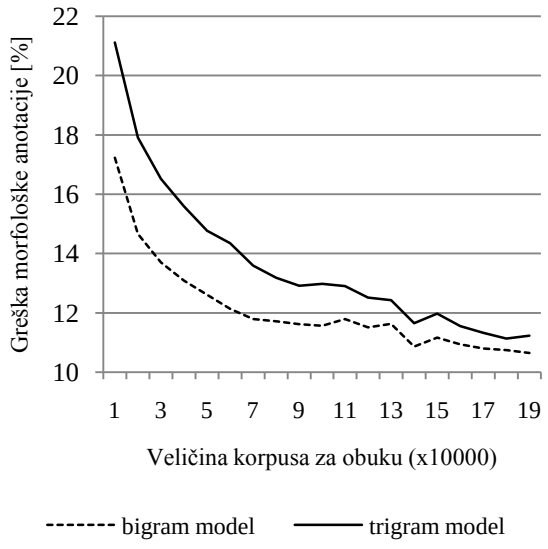
A. Eksperiment

Radi objektivne evaluacije rezultata svih eksperimenata na datom korpusu prethodno je utvrđeno da, ako se automatska morfološka anotacija vrši na slučajan način (slučajnim izborom jedne od mogućih oznaka za datu reč), vrednost greške anotacije iznosi 45,7% a greške akcentuacije 15,0%. Rezultate eksperimenata treba, dakle, evaluirati uzimajući ove vrednosti kao referentne, odnosno one u odnosu na koje je postignuto poboljšanje implementacijom određenog metoda (eng. *baseline*).

U okviru ovog istraživanja izvršeni su eksperimenti automatske morfološke anotacije na osnovu bigram i trigram modela. Radi utvrđivanja zavisnosti rezultata od veličine

korpusa za obuku, eksperimenti su vršeni na segmentima korpusa za obuku čija je veličina bila promenljiva u granicama od 10.000 do 190.000 u koracima od po 10.000, dok je veličina korpusa za testiranje u svakoj rundi eksperimenta bila konstantna i iznosila 10.000.

Rezultati eksperimenata prikazani su na slici 1. Trigram model se pokazuje konzistentno slabijim od bigram modela, što je bilo očekivano s obzirom na relativno skromnu veličinu AlfaNum tekstualnog korpusa za obuku i činjenicu da problem retkih podataka mnogo više dolazi do izražaja za viši red Markovljevog modela. Greška morfološke anotacije opada sa povećanjem korpusa za obuku i u oba slučaja se približava zasićenju. Slična tendencija može se zapaziti i kod greške akcentuacije, koja opada na 2,65% (bigram model), odnosno, 3,05% (trigram model).



Sl. 1. Zavisnost greške automatske morfološke anotacije od veličine korpusa za obuku

V. AUTOMATSKA MORFOLOŠKA ANOTACIJA NA OSNOVU TRANSFORMACIONIH PRAVILA

Morfološka anotacija zasnovana na primeni transformacionih pravila kao algoritam nadgledanog učenja predstavlja jedan od vrlo često korišćenih pristupa realizaciji automatskih sistema za anotaciju. Osnovna ideja na kojoj se ovaj pristup zasniva prvi put je izložena u [16], da bi u [17] bila detaljnije razrađena. Algoritam zasnovan na mašinskom učenju funkcioniše tako što sâm otkriva svoje slabosti i otklanja ih, time ujedno poboljšavajući svoje performanse.

Algoritam, naime, inicijalno dodeljuje svakoj reči određenu inicijalnu oznaku, dobijenu na osnovu analize obimnog korpusa za obuku ne vodeći računa o kontekstu. To po pravilu znači da se svakom tokenu dodeljuje oznaka koja se za taj token u korpusu za obuku najčešće javlja. Zatim se analizom istog korpusa, na osnovu određenih obrazaca za kreiranje transformacionih pravila, utvrđuje skup pravila koja transformišu jednu oznaku u drugu u odgovarajućem kontekstu i na taj način otklanjaju najveći broj grešaka inicijalne morfološke anotacije. Algoritam se zasniva na očekivanju da bi ista transformaciona pravila uklonila značajan broj grešaka inicijalne morfološke anotacije i kad bi se postupak sproveo na nekom drugom korpusu.

Ovakav pristup anotaciji prevazilazi dva najčešće pominjana nedostatka ekspertskih sistema – robustan je i praktično ne zahteva poznavanje gramatičkih pravilnosti datog jezika. U poređenju sa drugim automatskim metodama anotacije, poput statističkih, ističe se brojnim prednostima. Pored toga što je znatno manje zahtevan u pogledu potrebnog memorijskog prostora, identifikovana transformaciona pravila vrlo se jasno mogu tumačiti i vrednovati od strane čoveka, što se ne može reći za obimne tabele statističkih parametara do kojih dolaze statističke metode (uključujući Markovljeve modele). Zahvaljujući tome, identifikovana pravila se u slučaju potrebe vrlo jednostavno mogu dodatno modifikovati ručno, pa i koristiti za poboljšanje performansi ekspertskih sistema [18].

A. Osnovni algoritam

Brillov algoritam za automatsku morfološku anotaciju na osnovu transformacionih pravila obučava se u dva koraka. Prvi korak obuhvata inicijalnu anotaciju korpusa za obuku, a drugi identifikaciju transformacionih pravila.

U fazi inicijalne anotacije vrši se dodela inicijalnih oznaka svakom tokenu. S obzirom da je čitav korpus za obuku unapred morfološki anotiran, za svaki token je zapravo poznata i njegova prava oznaka, ali se taj podatak u ovoj fazi ne koristi. U originalnoj verziji algoritma bila je posmatrana relativna učestanost svake oznake za dati token u skupu za obuku i na osnovu toga su dodeljivane inicijalne oznake. Primera radi, reč *run* bila je inicijalno anotirana kao glagol u oba sledeća slučaja:

*We **run** three miles every day.* (1)

*The **run** lasted thirty minutes.* (2)

pri čemu je u primeru (1) to bila i ispravna anotacija, a u primeru (2) ne. Za zadovoljavajuće tačnu inicijalnu anotaciju reči koje se nisu javljale u korpusu za obuku bile su dovoljne dve jednostavne procedure koje su koristile informaciju o tome da li je reč napisana velikim početnim slovom kao i kojim trima slovima se ona završava.

Druga faza obuhvata dobijanje transformacionih pravila, koja se definišu na osnovu obrazaca kao što su sledeći:

Promeni oznaku t_i u oznaku t_j:

1. Ako prethodna (sledeća) reč ima oznaku t_k.
2. Ako reč dva mesta unazad (unapred) ima oznaku t_k.
3. Ako bilo koja od dve prethodne (sledeće) reči ima oznaku t_k.
4. Ako bilo koja od tri prethodne (sledeće) reči ima oznaku t_k.
5. Ako prethodna reč ima oznaku t_k, a sledeća reč ima oznaku t_l.
6. Ako prethodna (sledeća) reč ima oznaku t_k, a reč dva mesta unazad (unapred) ima oznaku t_l.
7. Ako je oblik prethodne (sledeće) reči L.
8. Ako je oblik prethodne (sledeće) reči L_1, a oblik reči dva mesta unazad (unapred) je L_2.

Transformaciono pravilo je identifikovano obrascem preko kog se konstruiše i konkretnom vrednošću t_i, t_j, t_k, t_l, L_1 odnosno L_2, to jest, svojim aktivacionim kontekstom (eng. *triggering environment*). Za svaki obrazac prolaskom kroz korpus za obuku identifikuju se pravila koja predstavljaju njegove instance i evidentira se broj grešaka koji bi primenom tih pravila bio ispravljen, kao i broj novih grešaka koji bi bio izazvan. Ilustracije radi, nakon inicijalne anotacije na osnovu segmenta Brown korpusa veličine 900.000 reči, za obrazac „Promeni oznaku t_i u oznaku t_j ako bilo koja od dve prethodne reči ima oznaku t_k“ pri prolasku kroz segment Brown korpusa veličine 50.000 reči utvrđeno je da pravilo konstruisano za (t_i, t_j, t_k) = (VB, NN, AT)[2] ispravlja 98 od ukupno 159 grešaka u inicijalno anotiranom korpusu, a unosi 18 novih [17]. Ukupna efikasnost određenog pravila na skupu za obuku može se izraziti razlikom između broja ispravljenih i broja unetih grešaka. U originalnoj verziji algoritma od svih transformacionih pravila uzima se najefikasnije i dodaje se u listu usvojenih transformacionih pravila, zatim se inicijalno anotirani korpus na osnovu njega reanotira i postupak akvizicije transformacionih pravila nastavlja se nad izmenjenim korpusom sve dok ne bude identifikovan unapred zadat broj transformacionih pravila ili dok efikasnost otkrivenih pravila ne padne ispod neke unapred zadate granice. Ovaj postupak naziva se *validacijom* transformacionih pravila.

Zanimljivo je da se analizom otkrivenih transformacionih pravila može videti da su sa lingvističke tačke gledišta najčešće sasvim logična, pogotovo ona koja su otkrivena među prvima. Primera radi, u engleskom jeziku, ukoliko se neka reč može anotirati i kao glagol u osnovnom obliku (VB) i kao imenica (NN), verovatnije je da će se raditi o imenici ako je neka od prethodne dve reči član (AT), što je ilustrovano primerom (1).

Kada su identifikovana sva transformaciona pravila, morfološka anotacija nad nepoznatim tekstom vrši se na isti način kao i nad tekstom za obuku. Prvo se izvrši inicijalna morfološka anotacija na osnovu učestalosti pojedinih oznaka u skupu za obuku, a zatim se sukcesivno primenjuju transformaciona pravila kako bi se broj grešaka morfološke anotacije smanjio. Transformaciono pravilo se primenjuje na reči koje se nalaze u kontekstu koji to pravilo definiše, ali samo u slučaju da je oznaka koju treba dodeliti nekoj reči zaista jedna od dozvoljenih oznaka koje ta reč može imati. U odsustvu morfološkog rečnika ovaj uslov bi se sveo na to da se transformaciono pravilo primenjuje samo onda ako reč već negde u korpusu za obuku ima oznaku koja bi trebalo da joj se dodeli. Ukoliko se neka reč nađe u kontekstu koji aktivira primenu dva ili više pravila, primenjuje se ono pravilo koje je na skupu za obuku imalo najveću efikasnost, odnosno, u najvećoj meri je umanjilo broj grešaka.

B. Modifikacije u cilju primene algoritma za anotaciju tekstova na inflektivnim jezicima

Zbog velike razlike u broju oznaka u Brown i AlfaNum korpusu, kao i u veličini samih korpusa, bilo je neophodno uvesti određene modifikacije u originalni Brillov algoritam kako bi mogao biti primenjen za automatsku morfološku anotaciju na srpskom jeziku.

Pre svega, Brillov algoritam umesto morfološkog rečnika koji bi poslužio kao izvor podataka za inicijalnu morfološku anotaciju koristi korpus od 900.000 reči. Ukoliko se token koji je potrebno inicijalno anotirati nalazi u ovom korpusu, inicijalno mu se dodeljuje ona oznaka sa kojom se on najčešće pojavljuje u korpusu. Zbog mnogo većeg broja različitih tokena u korpusu iste veličine na srpskom jeziku, čak i u korpusu od 900.000 reči bilo bi mnogo manje podataka na kojima bi trebalo estimirati raspodelu tokena po oznakama. Zato se inicijalna anotacija izvodi ne na osnovu najčešće oznake za datu reč, već se od svih dozvoljenih oznaka za datu reč bira ona koja se javlja najčešće u skupu za obuku, ali ne za datu reč, već uopšte. Primera radi, reč *knjiga*, koja može biti anotirana na dva načina, NNfs1--- i NNfp2---, biće inicijalno anotirana kao NNfs1---, ali ne zato što je u korpusu za obuku to češća oznaka za token *knjiga*, već zato što je to češća oznaka u korpusu za obuku uopšte.

U osnovnoj verziji algoritma transformaciona pravila usvajaju se jedno po jedno i korpus za obuku se pri usvajanju svakog pravila reanotira. S obzirom da je broj usvojenih transformacionih pravila u osnovnom algoritmu bio svega 71 i da se zbog razlika u veličini skupova za oznaku moglo očekivati da taj broj pri primeni odgovarajućeg algoritma na srpskom jeziku bude daleko veći (što je kasnije i potvrđeno rezultatima, koji pokazuju da se taj broj meri hiljadama), usvojena je nešto drugačija strategija akvizicije transformacionih pravila, koja se zasniva na vrednovanju pravila po grupama.

Za dalje poboljšanje performansi opisanog algoritma može se iskoristiti činjenica da, u slučaju inflektivnih jezika, specifična transformaciona pravila dobijena analizom korpusa vrlo često predstavljaju samo instance opštijih pravila. Primera radi, pravilo:

> *Ako je pridev praćen imenicom, promeni njegovu oznaku u genitiv množine ženskog roda pod uslovom da je i imenica ženskog roda i da je u genitivu množine;*

instanca je (hipotetičkog) opšteg pravila:

> *Ako je pridev praćen imenicom, promeni oznaku prideva tako da se vrednosti njegovog roda, broja i padeža podudare sa rodom, brojem i padežom te imenice.*

Strategija koja bi mogla da izvede opšta pravila na osnovu pojedinačnih bila bi od velike koristi za automatsku morfološku anotaciju, jer bi sistem mogao pravilno da postupi čak i u situacijama koje nisu eksplicitno prisutne u korpusu za obuku.

Formulisanje ovakve strategije olakšano je činjenicom da se koristi upravo pozicioni sistem oznaka. Prvi korak je grupisanje pravila prema obrascu od kog potiču, kao i vrsti reči koja odgovara polaznoj oznaci, ciljnoj oznaci i oznaci koja definiše aktivacioni kontekst. Primera radi, sva pravila koja nalažu da se oznaka prideva zameni oznakom prideva drugih vrednosti morfoloških kategorija ako je taj pridev praćen imenicom ($A \rightarrow A[N_{+1}]$) ispituju se zajedno,

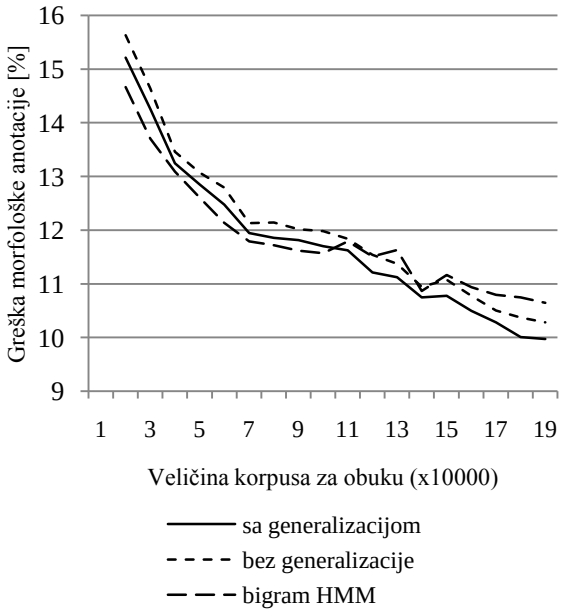
[2] VB = glagol (osnovni oblik), NN = imenica, AT = član.

da bi se ispitalo da li postoji mogućnost da su sva ta pravila, ili barem većina njih, instance jednog opšteg pravila. Detaljan prikaz ove modifikacije, kojom se obezbeđuje *generalizacija* transformacionih pravila, dat je u [14].

C. Eksperiment

U okviru ovog istraživanja izvršeni su eksperimenti automatske morfološke anotacije na osnovu transformacionih pravila, pri čemu je posebno ispitan uticaj predloženog metoda za generalizaciju transformacionih pravila. Kao i kod eksperimenata sa Markovljevim modelima, radi utvrđivanja zavisnosti rezultata od veličine korpusa za obuku, eksperimenti su vršeni na segmentima korpusa za obuku čija se veličina menjala u granicama od 20.000 do 190.000 u koracima od po 10.000, dok je veličina korpusa za testiranje u svakoj rundi eksperimenta iznosila 10.000. Deo korpusa za obuku veličine 10.000 u svakoj rundi eksperimenta korišćen je za validaciju pravila.

Rezultati eksperimenta prikazani su na slici 2. Zapaža se izvesno poboljšanje u pogledu tačnosti u odnosu na rezultate dobijene korišćenjem bigram HMM modela u slučajevima kada je korpus za obuku dovoljno velik. Postignut nivo greške morfološke anotacije od 9,97% predstavlja ujedno i najbolji rezultat dobijen u potpunosti automatskim sistemom za morfološku anotaciju.



Slika 2. Zavisnost greške anotacije od veličine korpusa za obuku (inicijalna anotacija na osnovu Markovljevih modela)

VI. Automatska morfološka anotacija korišćenjem ekspertskog sistema

A. AlfaNum ekspertski sistem

Razvoj ekspertskog sistema za automatsku morfološku anotaciju podrazumeva implementaciju već poznatih gramatičkih pravila, pri čemu se pojedini ekspertski sistemi razlikuju na osnovu formalizama za reprezentaciju gramatičkih pravila, kao i načina na koji ih interpretiraju. Iako upravo ekspertski sistemi za automatsku morfološku anotaciju ostvaruju najveću tačnost, njihov veliki nedostatak je u tome što zahtevaju stručno lingvističko znanje, a neretko i znatno veći trud u odnosu na onaj koji bi bio potreban za razvoj u potpunosti automatskog sistema. Pored toga, ovakvi sistemi su u velikoj meri zavisni od jezika, tako da modifikacija sistema radi primene na nekom čak i veoma srodnom jeziku može biti težak zadatak.

Rad na razvoju AlfaNum ekspertskog sistema za morfološku anotaciju [19] tekao je uporedo sa radom na razvoju morfološkog rečnika. Ovaj sistem koristi se kao komponenta u okviru AlfaNum sintetizatora na osnovu teksta čiji je glavni zadatak da pruži informacije potrebne za pravilnu akcentuaciju teksta, što je najvažniji preduslov za sintezu govora čija bi intonacija odgovarala prirodnoj.

Kao i opisani automatski sistemi, i ovaj polazi od rečenice rastavljene na tokene i oslanja se na morfološki rečnik da bi došao do skupa mogućih rečničkih unosa za svaku reč. Zadatak sistema je da otkrije niz rečničkih unosa koji u najvećoj meri poštuje unapred identifikovan skup gramatičkih pravila, pri čemu se pojam gramatičkog pravila ovde koristi u veoma širokom smislu, odnosno, pored klasičnih gramatičkih pravila obuhvata i bilo kakve druge pravilnosti u ponašanju pojedinih reči ili grupa reči. Pošto bi bilo nemoguće razmatrati sve moguće kombinacije rečničkih unosa posebno (u slučaju iole dužih rečenica njihov broj može biti neprihvatljivo velik), koristi se algoritam sličan dinamičkom programiranju, koji posmatra parcijalne hipoteze, odnosno nizove rečničkih unosa, držeći njihov broj pod kontrolom. Hipoteze koje u najmanjoj meri zadovoljavaju postojeći skup gramatičkih pravila (imaju najniži skor) odbacuju se, a hipoteza sa najvišim skorom se po završetku izvršavanja algoritma proglašava za važeću. Skor se u svakom koraku algoritma modifikuje na osnovu gramatičkog slaganja posmatrane reči sa rečima koje joj prethode, na osnovu obrazaca kao što su sledeći:

Uvećati skor parcijalne hipoteze $h = e_1,e_2,... e_m$ za n:

1. Ako unosu e_m odgovara oznaka t_i;
2. Ako unosu e_m odgovara oznaka t_i, a unosu e_{m-1} odgovara oznaka t_j;
3. Ako unosu e_m odgovara oznaka t_i, unosu e_{m-1} odgovara oznaka t_j, a unosu e_{m-2} odgovara oznaka t_k.

B. Eksperiment

Na nezavisnom korpusu veličine 10.000 reči utvrđeno je da je greška morfološke anotacije 6,75%, dok greška akcentuacije iznosi 1,26%. S obzirom da jedan isti token može biti pogrešno anotiran i u pogledu morfologije i u pogledu akcentuacije, treba pomenuti i da ukupna greška ekspertskog sistema (procenat unosa kod kojih se javlja bar jedna od ove dve greške) iznosi 6,95%.

VII. Zaključak

U ovom poglavlju dato je poređenje rezultata opisanih eksperimenata, odnosno uspešnosti različitih metoda automatske morfološke anotacije na srpskom jeziku. U tabeli 3 navedeni su uporedni rezultati svih eksperimenata, u pogledu greške morfološke anotacije i akcentuacije. Preglednosti radi, ukoliko je neki od eksperimenata izveden uz varijaciju određenih parametara, u tabeli je prikazan

samo najbolji postignut rezultat. Posmatrajući samo u potpunosti automatske tehnike, najveća tačnost morfološke anotacije postignuta je korišćenjem sistema koji se zasniva na primeni transformacionih pravila, a koristi inicijalnu morfološku anotaciju na osnovu bigram HMM modela i generalizaciju pravila kao način za prevazilaženje problema retkih podataka. Ovakav sistem postiže tačnost morfološke anotacije od približno 90%, što je u nivou rezultata koji se u literaturi navode za druge jezike sličnog stepena složenosti morfologije (87%-91%) [5], [20], pri čemu treba imati u vidu izrazitu zavisnost navedenih rezultata od jezika, vrste teksta i korišćenog skupa oznaka.

S druge strane, znatno veća tačnost morfološke anotacije, od preko 93%, postignuta je korišćenjem ekspertskog sistema koji se zasniva na direktnoj implementaciji gramatičkih pravila, uključujući i pravila koja opisuju specifično ponašanje pojedinih reči i grupa reči. Razlika u performansama ovog sistema u odnosu na potpuno automatske tehnike još je uočljivija u terminima greške akcentuacije. Najuspešnija automatska tehnika za akcentuaciju teksta (HMM, bigram model) ima nivo greške od 2,65%, dok ekspertski sistem ima nivo greške od svega 1,26%. Ovolika razlika delom je posledica i činjenice da je, imajući u vidu specifičnu namenu ekspertskog sistema u okviru sistema za sintezu govora, posebna pažnja u toku njegovog razvoja bila posvećena implementaciji gramatičkih pravila vezanih za akcentuaciju.

TABELA 3: POREĐENJE REZULTATA EKSPERIMENATA.

kratak opis	***greška morf. anotacije***	***greška akcentuacije***
HMM (bigram)	10,65%	2,65%
HMM (trigram)	11,23%	3,05%
transf. pravila, inicijalna anot. na osnovu relativnih učestalosti oznaka, sa generalizacijom pravila	10,42%	2,71%
transf. pravila, inicijalna anot. na osnovu HMM (bigram), sa generalizacijom pravila	9,97%	3,51%
ekspertski sistem	6,75%	1,26%

LITERATURA

[1] M. Sečujski, V. Delić, D. Pekar, R. Obradović, D. Knežević, "An overview of the AlfaNum text-to-speech synthesis system," in *Proc. SPECOM*, Moscow, Russia, 2007, pp.3-7. (Addenda Volume)

[2] C. Manning, H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press, 1993.

[3] W. N. Francis, H. Kučera, *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press, 1967.

[4] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, "Building a large annotated corpus of English: The Penn treebank," *Computational Linguistics*, No. 19, pp. 313–330, 1993.

[5] J. Hajič, "Building a syntactically annotated corpus: the Prague dependency treebank," *Issues of Valency and Meaning*, Karolinum, Prague, Czech Republic, pp. 106–132, 1998.

[6] J. Hajič, B. Hladká, "Czech language processing – POS tagging," in *Proc. 1st Intnl. Conf. on Language Resources and Evaluation*, Granada, Spain, 1998, pp. 931-936.

[7] L. van Guilder, "Automated part of speech tagging: a brief overview," *Handout for LING361*, Georgetown University, Georgetown, Washington DC.

[8] F. Karlsson, A. Voutilainen, J. Heikkilä, A. Anttila, *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Berlin, Germany: Mouton de Gruyter, 1995.

[9] C. Samuelsson, A. Voutilainen, "Comparing a linguistic and a stochastic tagger," in *Proc. 8th Conf. of the European Chapter of the Assoc. for Computational Linguistics*, Madrid, Spain, 1997, pp. 246-253.

[10] D. Hakkani-Tür, K. Oflazer, G. Tür, "Statistical morphological disambiguation for agglutinative languages," *J. of Computers and the Humanities*, 36(4), Kluwer Academic Publishers, The Netherlands, 2002, pp. 381-410.

[11] Đ. Kostić, *Kvantitativni opis strukture srpskog jezika: Korpus srpskog jezika*. Beograd, Srbija: Institut za eksperimentalnu fonetiku i patologiju govora, Filozofski fakultet, 2001.

[12] C. Krstev, D. Vitas, T. Erjavec, "MULTEXT-East resources for Serbian," in *Proc. Informational Society – Language Technologies Conf. IS-LTC*, Ljubljana, Slovenia, 2004, pp. 108-114.

[13] M. Sečujski, V. Delić, "A software tool for automatic part-of speech tagging in Serbian language," *Primenjena lingvistika*, 9(1), Beograd, Srbija: Društvo za primenjenu lingvistiku, 2008, pp. 97-103.

[14] M. Sečujski, "Automatska morfološka anotacija tekstova na srpskom jeziku," doktorska disertacija, Fakultet tehničkih nauka, Novi Sad, Srbija, 2009.

[15] A. Markov, " Пример статистического исследования над текстом 'Евгения Онегина', иллюстрирующих связь испытаний в цепь," (Primer statističkog istraživanja teksta „Evgenija Onjegina" kao ilustracija lančane povezanosti događaja) *Zbornik akademije nauka*, VI(7), St. Peterburg, Rusija, 1913, pp. 153-162.

[16] E. Brill, "A simple rule-based part of speech tagger," in *Proc. 3rd Conf. on Applied Natural Language Processing*, Trento, Italy, 1992, pp. 152-155.

[17] E. Brill, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," *Computational Linguistics*, 21(4), 1995, pp. 543-566.

[18] A. Kupusinac, M. Sečujski, "Povećanje tačnosti poluautomatske morfološke anotacije primenom transformacionih pravila," *TELFOR*, Beograd, Srbija, 2007, pp. 604-606.

[19] M. Sečujski, "Obtaining prosodic information from text in Serbian language," in *Proc. IEEE EUROCON*, Belgrade, Serbia, 2005, pp. 1654-1657.

[20] S. Džeroski, T. Erjavec, J. Zavrel, "Morphosyntactic tagging of Slovene: evaluating taggers and tagsets," in *Proc. 2nd Intnl. Conf. on Language Resources and Evaluation*, Athens, Greece, 2000, pp. 1099-1104.

ABSTRACT

The paper investigates the possibility of obtaining highly accurate part-of-speech (POS) tagging in Serbian language. The research included the design of a morphological dictionary of Serbian language, as well as a compatible POS tagged text corpus to be used for the experiments. Three approaches were implemented and compared: Hidden Markov models (HMMs), machine learning techniques based on transformation rules as well as an expert system. Influence of certain modifications of existing techniques on accuracy was also analysed. The best result was achieved when the expert system was used, and all obtained results are comparable to those obtained for other languages with a similar degree of complexity of morphology.

A COMPARISON OF TECHNIQUES FOR PART-OF-SPEECH TAGGING OF TEXTS IN THE SERBIAN LANGUAGE

Milan S. Sečujski

Aleksandar D. Kupusinac