SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Hrvoje Peradin

SINTAKTIČKA ANALIZA TEKSTOVA NA HRVATSKOM JEZIKU TEMELJENA NA GRAMATICI OGRANIČENJA

Diplomski rad

Voditelj rada:
doc. dr. sc. Jan Šnajder

Zagreb, 31.08.2012.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Obrada prirodnog jezika (engl. *natural language processing*, NLP) i računalna lingvistika (engl. *computational linguistics*) interdisciplinarna su područja koja objedinjuju različite grane računalne znanosti, logike i lingvistike.

Iako teorijski povijesni razvoj računalne obrade prirodnog jezika seže još do sedamnaestog stoljeća kod Descartesa i Leibniza koji su predlagali svojevrsni univerzalni *interlingua*, pravi razvoj područja započinje tek u dvadesetom stoljeću, već i prije značajnih napredaka u razvoju računalne tehnologije. Slikovit primjer je patent rusa P. Trojanskog [21], koji je 1933. predložio stroj za automatsko prevođenje na temelju Esperanta. Ipak, praktični razvoj sustava za NLP počeo je tek 1950. Od utjecajnijih publikacija iz tog razdoblja zanimljivo je izdvojiti Weaverov memorandum ([26],1949.) u kojem predlaže primjenu postojećih kriptografskih tehnika na automatsko prevođenje i Turingov članak *Computing Machinery and Intelligence* [24], u kojem se prvi put spominje danas poznati Turingov test. Nedugo zatim, IBM je 1954. proveo famozni eksperiment u Georgetownu [12] u kojem su demonstrirali naoko vrlo uspješan sustav za automatsko prevođenje ruski - engleski, na središnjem (engl. *mainframe*) računalu IBM 701. Iako je sustav imao samo šest gramatičkih pravila i vokabular od tek 250 riječi, djelovao je vrlo efektno i potaknuo vladu SAD-a na intenzivnije financiranje istraživanja računalne lingvistike. Sljedećih se deset godina u duhu Hladnog rata intenzivno radilo uglavnom na prevođenju između ruskog i engleskog jezika, no primjerice i na kineskom i japanskom. Glasoviti izvještaj ALPAC-a [14] 1966. godine na neko je vrijeme u SAD-u ugušio taj trend, zaključivši da su istraživanja neisplativa i da u području nema značajnih rezultata. Sljedeći značajni period obilježila su istraživanja na području umjetne inteligencije, baza znanja i logike. Zanimljiv primjer jest ELIZA Josepha Weizenbauma [27], program koji je vrlo uspješno simulirao rogerijanskog psihoterapeuta. Krajem osamdesetih, naglim rastom snage i brzine računala došlo je do procvata tehnika zasnovanih na strojnom učenju (engl. *machine learning*) i statističkoj obradi velikih korpusa tekstova. Za razliku od prethodnih tehnika koje su se zasnivale na velikoj količini pažljivo pisanih ručnih pravila, otvorile su se tehničke mogućnosti za automatsko ili poluautomatsko izvlačenje informacija iz samoga teksta. Razvojem interneta pojavljuju se i javno dostupni komercijalni alati za strojno prevođenje (primjerice

BabelFish[1], a kasnije i Googleovi jezični alati, oba zasnovana na Systranovom[2] softveru).

Ciljevi NLP-a, sežu dakle od konstrukcije isključivo teoretskih modela funkcioniranja prirodnog jezika do komercijalnih praktičnih primjena kao što su, primjerice, strojno prevođenje, interakcija između čovjeka i računala, različite vrste automatskog izvlačenja informacija iz teksta, generiranje prirodnog jezika, pretraživanje informacija i odgovaranje na pitanja. Pristupi rješavanju tih problema mogu se zasnivati na automatski generiranim modelima, na ručno pisanim pravilima ili pak na njihovoj kombinaciji. Automatski dobiveni modeli mogu se konstruirati u relativno kratkom vremenu, no za njih je potrebno imati na raspolaganju dovoljno velik skup za učenje, a čak i ukoliko nam je takav skup dostupan, takvi modeli imaju gornju teoretsku granicu kvalitete. S druge strane, modeli koji se zasnivaju na ručno pisanim pravilima manje ovise o količini sirovih resursa i njihova kvaliteta je u principu ograničena samo uloženim trudom, ali zahtijevaju stručno lingvističko znanje i dugotrajan razvoj.

Za razliku od većine umjetnih jezika koji su uglavnom domenski i egzaktni inherentno i najproblematičnije svojstvo prirodnog jezika je njegova višeznačnost. Kod morfološki bogatih jezika kao što je hrvatski morfološka višeznačnost uzrok je brojnim greškama već u početnoj fazi obrade teksta te je za ispravno funkcioniranje svakog sustava za obradu prirodnog jezika nužno točno označiti kategorije riječi. Gramatika ograničenja (engl. *Constraint Grammar*, CG) jest model zasnovan na ručno pisanim pravilima koji se može koristiti na više razina obrade jezika, a posebno je pogodan za razrješavanje morfosintaktičkih višeznačnosti.

U okviru ovog rada razvijen je skup pravila gramatike ograničenja za sintaktičku analizu tekstova na hrvatskom jeziku, zamišljen kao modul u cjevovodu za obradu tekstova na hrvatskom jeziku. Potpuna analiza obuhvatila bi morfosintaktičko označavanje, određivanje sintaktičke funkcije i analizu ovisnosnih odnosa riječi u rečenici. Ovaj rad usredotočen je na problem morfosintaktičkog označavanja, dok se za posljednja dva zadatka iznosi prijedlog daljnjeg razvoja. Pravila morfosintaktičkog označavanja razvijena su na osnovi korpusa novinskih tekstova lista *Vjesnik*, implementirana u programskom paketu `vislcg3`[3] i vrednovana na ručno označenom skupu od oko 5000 pojavnica.

Inspiracija za ovaj rad jest gramatika ograničenja razvijena na platformi za strojno prevođenje Apertium [15], u sklopu *open-source* implementacije jezičnog para srpsko-hrvatski – makedonski (apertium-sh-mk)[4] koji je razvijen za vrijeme Google Summer of Code 2011.[5] Na osnovi iskustva stečenog u radu s Apertiumom u potpunosti je iznova razvijena gramatika ograničenja za razrješavanje morfosintaktičkih višeznačnosti. Nova gramatika je opširnija i bolje strukturirana, koristi detaljniji skup oznaka i drugačiji mor-

[1] `http://http://www.babelfish.com/`
[2] `http://www.systran.co.uk`
[3] `http://beta.visl.sdu.dk/cg3.html`
[4] `http://wiki.apertium.org/wiki/Serbo-Croatian_and_Macedonian`
[5] `http://code.google.com/soc/`

fološki analizator te je usredotočena na hrvatski jezik.

Rad je organiziran u pet poglavlja. U prvom je poglavlju dan podrobniji opis postupaka morfosintaktičkog označavanja i gramatike ograničenja te je dan pregled srodnih radova. U drugom poglavlju opisan je razvoj gramatike ograničenja za morfosintaktičko označavanje riječi u hrvatskom jeziku. U trećem poglavlju nalazi se prijedlog proširenja sustava za punu sintaktičku analizu hrvatskog jezika. Zaključak rada nalazi se u petom poglavlju. Četvrto poglavlje sadrži eksperimentalno vrednovanje dobivenog sustava i analizu njegovih prednosti i nedostataka.

U rad su uključena i dva dodatka. Dodatak A sadrži prikaz djelovanja nekih pravila označavanja na korpusu. U dodatku B je isječak iz ručno označenog korpusa.

1 Gramatika ograničenja

1.1 Označavanje vrste riječi

Označavanje vrste riječi (engl. *tagging*) važan je korak kod većine postupaka obrade prirodnog jezika. To je postupak u kojem za ulazni niz tokena svakome od njih pridružujemo neku oznaku (engl. *tag*). Oznake mogu odgovarati različitim razinama obrade jezika. Na osnovnoj razini obično se radi s oznakama vrste riječi (engl. *part of speech tags, POS tags*) ili s morfosintaktičkim oznakama.

U obradi jezika riječi se obično radi lakše kategorizacije i ispravne obrade na višim jezičnim razinama lematiziraju. Tako razlikujemo leme i pojavnice (engl. *word forms*). Pojavnica je puni oblik riječi koji se pojavljuje u tekstu, a lema je obično neki kanonski ili rječnički oblik, kod glagola primjerice infinitiv:

$$\textit{radim}, \textit{radiš}, \textit{radi} \rightarrow \textit{raditi}.$$

Za svaku pojavnicu njezinu lemu s pridruženim morfološkim oznakama zovemo *čitanje* (engl. *reading*), npr.:

$$\textit{ženama} \equiv \textit{žena}\textbf{.n.f.pl.dat}.$$

Često se jednoj pojavnici može pridružiti više čitanja (ponekad i s više različitih lema). Skup svih čitanja za zadanu pojavnicu zovemo *kohorta* (engl. *cohort*). Primjerice

$$\begin{aligned} \text{među} &\equiv \text{među}\textbf{.pr.ins} \\ &\equiv \text{među}\textbf{.pr.acc} \\ &\equiv \text{međa}\textbf{.n.f.sg.acc} \end{aligned}$$

Višeznačnost prirodnog jezika jedno je od njegovih najnezgodnijih svojstava, i kod većine sustava predstavlja zapreku već u početnoj fazi obrade. Ta se višeznačnost ispoljava na svim razinama jezika: na morfološkoj, sintaktičkoj i semantičkoj. U kontekstu morfosintaktičkog označavanja rješava se problem višeznačnosti na morfološkoj razini.

Oznaka	Značenje
n	imenica
v	glagol
prn	zamjenica
det	član
pr	zamjenica

Tablica 1.1: Primjer oznaka vrsta riječi (engl. *POS tags*)

Kod morfološki jednostavnijih jezika, kao što su npr. španjolski ili engleski, uglavnom je dovoljno promatrati višeznačnosti na razini vrste riječi (engl. *part-of-speech ambiguity*).

Primjerice, uz jednostavan skup oznaka kao u tablici 1.1 španjolskoj rečenici

Vino a la playa

možemo pridružiti sljedeće oznake:

Vino**.n** a**.pr** la**.det** playa**.n** ≡ Vino na plažu!
Vino**.v** a**.pr** la**.det** playa**.n** ≡ Došla je na plažu.
Vino**.n** a**.pr** la**.prn** playa**.n** ≡ Vino joj, plažo (??)
Vino**.v** a**.pr** la**.prn** playa**.n** ≡ Došla joj je, plažo(?)

Česta višeznačnost vrste riječi u engleskom jest ona između imenice i glagola:

book**.v** the seats ≡ rezervirati mjesta
read a book**.n** ≡ pročitati knjigu

Kod morfološki složenih jezika kompliciraju se i višeznačnosti. Česta višeznačnost u hrvatskom jeziku jest između genitiva jednine i množine.

Na ulici stoji žena**.n.gen.sg** (jednina)
Na ulici ima mnogo žena**.n.gen.pl** (množina)

Dativ, lokativ i instrumental množine u hrvatskom jeziku uvijek su identični:

Rukama **(čime?)** sam zagrabio pijesak.
Rukama **(čemu?)** sam zagrabio pijesak.
(u) Rukama **(u čemu?)** sam zagrabio pijesak.

ili

Ljudima **(kome?)** sam pomeo pod.
Ljudima **(kime?)** sam pomeo pod.

Posljednja dva primjera ilustracija su semantičke višeznačnosti, kod koje značenje riječi ovisi o širem kontekstu ili izvanjezičnome znanju.

Neki uobičajeni pristupi označavanju vrsta riječi

Skriveni Markovljevi modeli – HMM

Označivači zasnovani na skrivenom Markovljevom modelu (engl. *hidden Markov model, HMM*) rade na principu *n*-grama.[1] Označivač se izgrađuje tako da se na nizovima od uzastopnih *n* tokena metodama nadziranog strojnog učenja trenira skriveni Markovljev model. Vjerojatnosti modela mogu se na temelju frekvencija *n*-grama u označenom korpusu procijeniti metodom najveće izglednosti (engl. *maximum likelihood estimation, MLE*). Takav model daje vjerojatnosti uzastopnog pojavljivanja *n* pojavnica, tako da je svakoj pridružena jednoznačna oznaka.

Primjerice, za 2-gram (bigramski) označivač, uz skup oznaka

$$\Gamma := \{\gamma_1, \gamma_2, \ldots, \gamma_{|\Gamma|}\} = \{n, vb, adj...\}$$

i klase višeznačnosti

$$\Sigma := \{\sigma_1, \sigma_2, \ldots, \sigma_{|\Sigma|}\} = \{n|vb, det|prn...\}$$

skriveni Markovljev model M može se definirati kao:

$$M := (A, B, \pi)$$

gdje je A matrica tranzicijskih vjerojatnosti

$$a_{ij} = p(\gamma_j|\gamma_i)^1$$

a B matrica emisijskih vjerojatnosti

$$b_{ij} = p(\sigma_j|\gamma_i)^2$$

U vektoru $\pi = (\pi_1, \ldots, \pi_{|\Gamma|})$ nalaze se početne vjerojatnosti; tj. element π_i je vjerojatnost da se γ_i nalazi na početku rečenice.

[1] bilo koji niz uzastopnih pojavnica duljine n; npr. unigram, bigram, ...

[1] vjerojatnost da se oznaka γ_j pojavljuje nakon oznake γ_i,

[2] vjerojatnost da se klasa višeznačnosti σ_j pojavljuje nakon oznake γ_i, u neoznačenom korpusu

Uz takav izgrađen model neviđeni se tekst zatim označava Viterbijevim algoritmom[2], ili nekom drugom metodom dinamičkog programiranja.

HMM označivači obično su iznimno fleksibilni i robustni i dostižu točnosti od oko 96%.

Brillov označivač

Brillov označivač[10]. jedan je od najpoznatijih označivača zasnovanih na pravilima. Označivač radi u dva koraka. U prvom koraku svakoj riječi dodjeljuje najvjerojatniju oznaku, zasnovanu na automatskoj procjeni temeljem velikog, označenog korpusa.

U drugom koraku označivač koristi dva trika kako bi popravio oznake dodijeljene u prvome koraku. Prvo, pretpostavlja da su nepoznate riječi koje počinju velikim slovom vlastite imenice i dodjeljuje im tu oznaku. Zatim pokušava označiti preostale nepoznate riječi istom onom oznakom kojom je označena neka riječ u korpusu za učenje, a koja završava na identična tri slova kao i nepoznata riječ (npr. *copious*, *erroneous*). Označivač zatim primjenjuje automatske "zakrpe" na pravila, koje se utvrđuju na temelju korpusa uporabom sljedećih predložaka:

- riječ ima oznaku **tag1** i nalazi se u kontekstu **C** $\Rightarrow$ promijeni oznaku u **tag2**
- riječ ima oznaku **tag1** i ima leksičko svojstvo **P** $\Rightarrow$ promijeni oznaku u **tag2**
- riječ ima oznaku **tag1** i nalazi se u regiji **R** $\Rightarrow$ promijeni oznaku u **tag2**

Zatim se po još opširnijem skupu predložaka traže "zakrpe" koje minimiziraju grešku. Primjerice, predložak

```
TO IN NEXT-TAG AT
```

Mijenja oznaku `TO` (infinitiv) u `IN` (prijedlog) ukoliko je sljedeća oznaka `AT` (član), dok predložak

```
NP NN CURRENT-WORD-IS-CAP NO
```

mijenja vlastitu imenicu u opću imenicu ukoliko ona ne počinje velikim slovom. Točnost koju doseže Brillov označivač obično iznosi oko 95%.

Neki standardni klasifikatori iz strojnog učenja također se uspješno primijenjuju na označavanje riječi, primjerice *SVM*(engl. *support vector machine*), *maximum-entrophy*) i *k-NN*. Svi ovi klasifikatori uglavnom dosežu točnost od oko 95%.

[2] Viterbijev algoritam za zadani niz računa njegov prolazak kroz Markovljev model, takav da taj prolazak ima maksimalnu vjerojatnost.

1.2 Gramatika ograničenja

Gramatika ograničenja (engl. *Constraint Grammar*, CG) metodološka je paradigma za obradu prirodnog jezika koja predstavlja izravan napad na jezične višeznačnosti [19]. Ručno pisana pravila koriste se za ograničavanje mogućih gramatičkih čitanja riječi ovisno o kontekstu u kojem se riječ pojavljuje, čime se smanjuje višeznačnost riječi. Čitanja riječi mogu biti morfološka, sintaktička ili semantička, i mogu se koristiti na različitim razinama jezične obrade, od morfološkog označavanja i parsanja do semantičkog označavanja, razrješavanja anafore, izvlačenja informacija i strojnog prevođenja.

Performanse zrelih označivača temeljenih na gramatici ograničenja iznimno su visoke, s kvalitetom izraženom u F-mjeri od oko 99% za *POS* oznake i 95% za označavanje sintaktičkih funkcija riječi. Za razliku od stohastičkih označivača, kod označivača temeljenih na gramatici ograničenja ne postoji teoretska gornja granica performansi, jer je točnost uvijek moguće povećati dodavanjem novih pravila.

Formalizam gramatike ograničenja sastoji se od niza operacija s kontekstnim testovima, koja se iterativno primjenjuju na pojavnice u tekstu.

Primjerice, na rečenicu:

Vino**.v** a**.pr** la**.det/prn** playa**.n**

mogli bismo primijeniti pravilo:

```
SELECT (det) IF (0 (det) OR (prn)) (1 (n) )
```

Ukoliko je uvjet desno od `IF` zadovoljen (tj. ukoliko je riječ o višeznačnosti **det/prn** nakon koje slijedi **n**) gramatika će izvršiti operaciju odabira oznake **det**, i time ukloniti sva čitanja koja tu oznaku ne sadrže.

Brojevi označavaju relativan pomak u odnosu na pojavnicu koja se trenutno obrađuje (`0` označava trenutnu pojavnicu, `1` jedno mjesto udesno, `-1` mjesto ulijevo, itd.)

Gramatika ograničenja iznimno je fleksibilna po pitanju oznaka, koje praktički mogu biti proizvoljne te dozvoljava njihovo grupiranje u gramatičke kategorije unutar preambule gramatike, ključnom riječi `LIST`.

```
LIST Noun = n np ;
LIST Adjective = adj ;
LIST Verb = vblex vbcop vbhaver vbmod vbaux ;
LIST Pronoun = prn ;
LIST Numeral = num ;
LIST Ordinal = ord;
LIST Cardinal = crd ;
```

Na tako definiranim kategorijama moguće je pomoću ključne riječi SET skupovnim operacijama definirati njihove unije (oznaka |) ili razlike (oznaka -). Primjerice, za kategoriju imenskih riječi i kardinalnih brojeva:

```
SET CardNum  = Numeral - Ordinal ;
SET Nominal = Noun | Pronoun | Adjective | Numeral | CardNum;
```

Na skupovima je moguće izvoditi i druge standardne operacije, koje nećemo posebno opisivati. Navodimo samo primjer Kartezijevog produkta (oznaka +) koji se često pojavljuje u pravilima. Primjerice, ovako se može izraziti imenicu ženskog roda u nominativu jednine:

```
LIST Noun = n np ;
LIST Nominative = nom ;
LIST Feminine = f ;
LIST Singular = sg ;

SET Imenica = Noun + Feminine + Singular + Nominative ;
```

Pravila se primjenjuju u odsječcima, nad nekim unaprijed definiranim prozorom teksta. Prozor teksta obično je rečenica, i definira se preko delimitera koji se također navode u prembuli.

```
DELIMITERS = "<.>" "<!>" "<?>" "<...>" "<¶>" ;
```

Pravila se primjenjuju po principu:

```
za svaki (prozor):
  za svako (pravilo):
    za svaku (kohortu):
      primijeni(pravilo);
```

Pravila najčešće obavljaju odabir (SELECT), uklanjanje(REMOVE) ili zamjenu (SUBSTITUTE) čitanja.

Ovaj se postupak ponavlja sve dok na kohortu više nije moguće primijeniti niti jedno pravilo, ili je na kohorti preostalo jedno jedino čitanje. Gramatika ograničenja nikad neće ukloniti zadnje preostalo čitanje, čime se povećava robustnost sustava, jer se dopušta da postupak označavanja prežive nekonvencionalne oznake.[3]

[3]Primjerice, oznake koje kod stohastičkog označivača ne bi nikad bile pridijeljene zbog niske vjerojatnosti.

1.3 Srodni radovi

Morfosintaktičko označavanje za hrvatski jezik

Prijašnji radovi na morfosintaktičkom označavanju hrvatskih tekstova bili su fokusirani na stohastičke označivače [3, 20]. U [3] opisano je treniranje i vrednovanje označivača TnT[9], temeljenog na modelu HMM, na korpusu *Croatia Weekly* od 100.000 riječi. Korpus je bio ručno označen, korištenjem oko 1000 različitih morfosintaktičkih oznaka (tzv. morfosintaktičkih deskriptora, MSD-ova) prema normi MULTEXT-East v3 [13]. Točnost označivača ispitana je na nasumce odabranih 10% korpusa, a ostalih 90% je korišteno za treniranje modela označivača. Dobiveni rezultati za oznake vrste riječi u prosjeku su iznosili iznad 98% točnosti, što dostiže razinu ljudske pogreške, dok je za pune morfosinaktičke oznake (MSD-ove) greška varirala između 86%-95%. Autori zaključuju kako je do većine grešaka kod MSD-označavanja dolazilo na vrstama riječi s najvećim brojem mogućih MSD-ova – imenicama, zamjenicama i pridjevima, dok su greške u određivanju vrste riječi bile gotovo zanemarive.

U [20] implementiran je označivač zasnovan na modelu HMM. Označivač je treniran i razvijen na korpusu nasumce odabranih novinskih članaka iz lista *Vjesnik*. Korpus je sadržavao 20.000 pojavnica, i bio ručno označen posebno razvijenim skupom morfosintaktičkih oznaka, koji je oblikovan na osnovi kategorija vrsta riječi prema normi MULTEXT-East. Model HMM treniran je na nasumce odabranom skupu za učenje od 15.000 tokena, dok je ostatak skupa uzet za ispitivanje. Označivač je postizao točnost od 83,64% na nepoznatim riječima, a ukupna točnost mu je iznosila 92,33%.

Vrijedi spomenuti da slični radovi sa sličnim rezultatima postoje i za označavanje tekstova na srpskom jeziku, primjerice [11]. Ovdje je riječ o označivaču zasnovanom na transformacijama razvijenom na korpusu od oko 200.000 tokena, čiji je postotak greške iznosio 10%.

Označivač razvijen u okviru ovog diplomskog rada zamišljen je prvenstveno kao dopuna postojećim sustavima.

Gramatika ograničenja za strane jezike

Formalizam gramatike ograničenja uspješno je primjenjen na brojne jezike, uključujući engleski [19, 8], baskijski [1], portugalski [4], francuski [6], danski [5], norveški [17] i španjolski [7].

U tablici 1.3 dan je usporedni prikaz gramatika ograničenja za druge jezike, zajedno s leksikonima koje imaju na raspolaganju. Većina gramatika ne funkcionira samo na morfosintaktičkoj razini, već sadrže i pravila za određivanje semantičkih uloga, pravila za gramatiku ovisnosti i pravila gramatike fraznih struktura. Više detalja može se naći na `http://beta.visl.sdu.dk/constraint_grammar_languages.html`.

Tablica 1.2: Usporedni prikaz gramatika ograničenja za druge jezike

Jezik	Parser	Leksikon	Gramatika
Danski	*DanGram*	100.000 riječi, 40.000 naziva	8.000 pravila
Portugalski	*Palavras*	70.000 riječi, 15.000 naziva	7500 pravila
Španjolski	*HISPAL*	73.000 riječi	4500 pravila
Engleski	*EngCG*	180.000 riječi	4400 pravila
Francuski	*FrAG*	180.000 riječi	2020 pravila
Njemački	*GerGram*	232.000 riječi	1940 pravila
Esperanto	*EspGram*	32.000 riječi	2600 pravila
Talijanski	*ItaGram*	72.000 riječi	3290 pravila
Švedski	*SveGram*	63.000 riječi	prilagođena danska

U okviru platforme za strojno prevođenje Apertium [15] na osnovi gramatike ograničenja u posljednje su vrijeme razvijeni alati i za neke slabije zastupljene jezike, za koje postoje ograničeni resursi (turski, islandski, bretonski, velški, makedonski, ruski i "srpskohrvatski"[4]).

[4]Radi se implementaciji koja primjenom odgovarajućih pravila cilja obuhvatiti oba jezika.

2 Gramatika ograničenja za označavanje vrste riječi na hrvatskom jeziku

2.1 Skup oznaka

Raniji radovi na označavanju riječi na hrvatskom jeziku [2, 3] koriste skup oznaka zasnovan na morfosintaktičkim deskriptorima (MSD-ovima) prema normi MULTEXT-East [13]. MULTEXT-East sadrži opširne i usklađene leksičke specifikacije za devet istočnoeuropskih jezika, uključujući hrvatski, što ga čini vrlo interoperabilnim i pogodnim za standardizaciju. Unatoč širokoj prihvaćenosti norme MULTEXT-East, ona nije korištena u ovom radu. Glavni razlog za to je što MULTEXT-East koristi pozicijske oznake, koje su nepraktične za upotrebu s gramatikom ograničenja. Iz tog razloga razvijen je nov, nepozicijski skup oznaka, prikazan tablicom 2.1.

Skup je temeljen na oznakama korištenima u apertium-sh-mk, no proširen je kako bi pokrio većinu MSD-ova za hrvatski jezik. U nekim sastavnicama skup je opširniji od onog definiranog normom MULTEXT-East. Primjerice, u skup su uključene oznake za tranzitivnost i glagolski vid, posebne oznake za futur prvi i drugi, kao i oznake za interpunkcije. Skup također sadrži dodatne oznake za semantičke klase imenica, koje bi mogle biti korisne za kasniju obradu na semantičkim razinama. Kako je jezgra skupa oznaka u principu nepozicijska varijanta MULTEXT-Easta, bez problema ga je moguće pretvoriti u pozicijske oznake pa ovim proširenjem nije izgubljena kompatibilnost.

Tablica 2.1: Morfološke oznake za hrvatski jezik

POS tag	Value tags
abbr (skraćenica)	**n** (imenička), **adv** (priložna), **adj** (pridjevska) **spl** (jednostavna), **cpx** (složena), **ant** (ime), **cog** (prezime), **top** (toponim), **alt** (ostalo), **sg** (jednina), **pl** (množina) + **PADEŽ** + **ROD**
adv (prilog)	**pst** (pozitiv), **cmp** (komparativ), **sup** (superlativ)
adj (pridjev)	**qlf** (opisni), **pos** (posvojni), **pst** (pozitiv), **cmp** (komparativ), **sup** (superlativ), **sg**, **pl**, **ind** (neodređeni), **def** (određeni) + **CASE** + **GEND**
cnj (veznik)	**coo** (nezavisni), **sub** (zavisni), **spl** (jednostavan), **cpx** (složen)
ij (usklik)	**spl** (jednostavan), **cpx** (složen)
n (imenica)	**com** (opća), **prp** (vlastita), **ant** (ime), **cog** (prezime), **top** (toponim), **alt** (drugo), **sg**, **pl** + **PADEŽ** + **ROD**
num (broj)	**crd** (kardinalni), **ord** (redni), **mlt** (višestruki), **dgt** (znamenka), **rom** (rimski), **ltr** (slovima) + **PADEŽ** + **ROD**
part (čestica)	**neg** (niječna), **int** (upitna), **mod** (modalna), **aff** (potvrdna)
pr (prijedlog)	**spl** (jednostavan), **cpx** (složen) + **PADEŽ**
prn (zamjenica)	**prs** (osobna), **dem** (pokazna), **ind** (neodređena), **pos** (posvojna), **int** (upitna), **rel** (odnosna), **rfx** (povratna) **p1** (1. lice), **p2** (2. lice), **p3** (3. lice), **sg**, **pl**, **clt** (klitika) + **PADEŽ** + **ROD**
vb (glagol)	**lex** (leksički), **aux** (pomoćni), **mod** (modalni), **cop** (kopula) **ind** (indikativ), **imp** (imperativ), **cnd** (kondicional), **inf** (infinitiv), **pp** (particip), **prs** (prezent), **ipf** (imperfekt), **aor** (aorist), **futI** (futur I), **futII** (futur II), **pct** (perfekt), **pqp** (pluskvamperfekt) **p1** (1. lice), **p2** (2. lice), **p3** (3. lice) **neg** (niječni) **perf** (svršeni) **imperf** (nesvršeni) **tv** (prijelazni) **iv** (neprelazni) **act** (aktiv), **psv** (pasiv)
(interpunkcija)	**qt** (navodnici), **aps** (apostrof), **lqt** (lijevi navodnik), **rqt** (desni navodnik), **bqt** (početak citata), **eqt** (kraj citata), **cm** (zarez), **sep** (odvajajući), **cnt** (nastavljajući), **lpar** (lijeva zagrada), **rpar** (desna zagrada), **sent** (kraj rečenice), **fxq** (točka, uskličnik, upitnik) **hyp** (spojnica), **dsh** (povlaka) **pct** (ostalo)

PADEŽ = **nom**, **gen**, **dat**, **acc**, **voc**, **loc**, **ins**
ROD = **m** (muški), **mi** (muški neživo), **ma** (muški živo), **nt** (srednji), **f** (ženski)

2.2 Struktura sustava

Označivač na osnovi gramatike ograničenja obično radi u tri koraka:

- morfološka analiza,
- razrješavanje morfoloških višeznačnosti,
- dodjeljivanje sintaktičkih oznaka.

U prvome se koraku na ulaznu pojavnicu primjenjuje morfološki analizator, čime se dobiva skup kohorti.

Drugi je korak primjena pravila gramatike. Idealno, ona za svaku kohortu daju jedinstveno i točno čitanje. Pravila gramatike mogu isključiti netočnu analizu, odabrati točnu ili pojavnicama dodijeliti nove oznake, ovisno o kontekstu u kojem se pojavljuju.

Primjer 2.1: Primjer konverzije kohorte za pojavnicu 'da' u nepozicijske oznake

```
da        da:Css da:Qr dati:Vmia2s dati:Vmia3s dati:Vmip3
                              ⟶
        "<da>"
                "da" cnj sub
                "da" part aff
                "dati" vblex aor p2 sg tv perf
                "dati" vblex aor p2 sg perf iv
                "dati" vblex aor p2 sg ref perf
                "dati" vblex aor p3 sg tv perf
                "dati" vblex aor p3 sg perf iv
                "dati" vblex aor p3 sg ref perf
                "dati" vblex pres p3 sg tv perf
                "dati" vblex pres p3 sg perf iv
                "dati" vblex pres p3 sg ref perf
```

U trećem se koraku pojavnicama koje su posve morfološki razriješene dodjeljuju sintaktičke uloge. Za to je potrebna mnogo opširnija gramatika no što ju je bilo moguće razviti u okvirima ovog rada pa je dodjeljivanje sintaktičkih uloga tek načelno razmotreno u trećem poglavlju.

Morfološki analizator

Morfološku analizu obavlja flektivni leksikon iz [25]. Leksikon je dobiven poluautomatskom akvizicijom iz neoznačenog korpusa i trenutno pokriva 66.500 lema i 3.5 milijuna pojavnica, s ukupno 318 jedinstvenih oznaka. Broj jedinstvenih oznaka koje se pojavljuju u našem zlatnom standardu je 487, a teoretska gornja granica našeg skupa oznaka je 6200. Za svaku pojavnicu u tekstu leksikon daje kohortu s lematiziranim analizama i MSD-ovima prema normi MULTEXT-East.

Prije no što se kohorte može obraditi pravilima gramatike skup MSD-ova preslikava se u naš skup oznaka (2.1). Preslikavanje nije injektivno. U nekim se slučajevima MSD-ovi spajaju u jednu oznaku (npr. oznake za muški rod živo `ma` i muški rod neživo `mi`), a u nekim slučajevima se oznake izostavljaju (npr. indikativ ili aktiv za glagole).

Skup MSD-ova ne sadrži oznake za refleksivnost, tranzitivnost i perfektivnost. Kako su u mnogim slučajevima ove oznake vrlo korisne ili čak nužne za puno morfološko razrješavanje, MSD-ove smo tamo gdje je bilo moguće nadopunili iz dodatnih resursa. Koristili smo valencijski leksikon hrvatskih glagola (CROVALLEX, [22]) i morfološki analiza-

tor iz apertium-sh-mk. Ovi jezični resursi zajedno pokrivaju oko 2400 različitih glagolskih lema.

Pravila gramatike

Ostvarena gramatika sadrži 290 pravila za razrješavanje višeznačnost, što je više od polazne gramatike iz apertium-sh-mk (trenutno 170 pravila), no ipak mnogo manje nego gramatike zrelih sustava, kao što je primjerice [8] (gotovo 1.500 pravila). Pravila su implementirana u formalizmu CG3 i prevedena *open-source* prevodiocem `vislcg3`.[1]

Za razvijanje pravila koristili smo korpus od 1M pojavnica sastavljen od rečenica iz novinskih članaka *Vjesnika*.

Gramatika je strukturirana u tri dijela:

- čišćenje
- sigurna pravila
- heuristike

U prvome se dijelu izlaz analizatora čisti i modificira kako bi se izlazom moglo spretnije baratati. Neka se čitanja sažimlju, dok se nekima dodjeljuju preliminarne sintaktičke oznake, koje se kasnije koriste za brže uklanjanje netočnih čitanja. Ovdje se preliminarno primjenjuju i heuristike, primjerice za određivanje semantičke kategorije vlastite imenice.

U drugome se dijelu primjenjuju tzv. *sigurna pravila*, koja podrazumijevaju uklanjanje gramatičkih besmislica, ali ostavljaju mjesta semantičkim višeznačnostima. Ovaj skup pravila osmišljen je kao jezično univerzalan skup koji se može primijeniti na bilo kojem jezičnom žanru.

Treći dio čine heuristike, koje nisu toliko striktne i odražavaju lingvističku intuiciju. Sadrže tip pravila koji će najbolje funkcionirati na žanru kojem pripada tip teksta na kojem su pravila razvijana.

2.3 Primjeri pravila

Slijede primjeri zanimljivijih pravila za razrješavanje čestih morfoloških višeznačnosti u hrvatskom jeziku. Detaljniji prikaz ovih pravila zajedno s primjerima njihovog djelovanja na korpus nalazi se u Dodatku A. Za podrobniji opis sintakse pravila čitatelja upućujemo na dokumentaciju sustava `vislcg3`.[1]

[1] `http://beta.visl.sdu.dk/cg3.html`

Pridjev ili prilog

U hrvatskom se jeziku svakom pridjevu u nominativu jednine srednjeg roda može dodijeliti pridjevsko ili priložno čitanje. Ovaj kontrast ilustriraju rečenice:

(1) *Dobro dijete.*

(2) *Dobro izgledaš.*

Za prvu rečenicu morfološki analizator daje:

```
"<Dobro>"
        "dobar" adj nt sg nom ind
        "dobar" adj nt sg nom def
        "dobar" adj nt sg acc ind
        "dobar" adj nt sg acc def
        "dobar" adj nt sg voc ind
        "dobar" adj nt sg voc def
        "dobro" adv
"<dijete>"
        "dijete" n nt sg nom
        "dijete" n nt sg acc
        "dijete" n nt sg voc
```

A za drugu:

```
"<Dobro>"
       "dobar" adj nt sg nom ind
       "dobar" adj nt sg nom def
       "dobar" adj nt sg acc ind
       "dobar" adj nt sg acc def
       "dobar" adj nt sg voc ind
       "dobar" adj nt sg voc def
       "dobro" adv
"<izgledaš>"
       "izgledati" vblex iv prs p2 sg
```

U rečenici 1, *dobro* je pridjev u nominativu srednjeg roda , dok u (2) ima funkciju priloga. Točno čitanje za riječ *dijete* je imenica u nominativu jednine, a riječ *izgledaš* je

prezent glagola *izgledati* u drugom licu jednine. Dakle, ako iza riječi slijedi imenica u srednjem rodu, treba odabrati pridjevsko čitanje, a ako je riječ koja slijedi glagol, odabiremo prilog.

Pravila kojima se ovo postiže su:

```
SELECT Adjective IF (0 Adverb OR Adjective)(1 Noun + Neuter)
SELECT Adverb IF (0 Adverb OR Adjective)(1 Verb)
```

Akuzativ i tranzitivnost

Uz imeničku kategoriju roda hrvatski jezik za imenske riječi u muškom rodu jednine sadrži i kategoriju živosti. Živost se ispoljava u jednakosti akuzativa i nominativa:

Gledam kroz prozor**.nom/acc**

ili u jednakosti genitiva i akuzativa:

Gledam u čovjeka**.gen/acc**

Stoga su kod imenskih riječi u muškom rodu jednine višeznačnosti nominativ/akuzativ ili genitiv/akuzativ vrlo česte.

Kod ove višeznačnosti izrazito je korisno imati na raspolaganju informaciju o tranzitivnosti glagola, bez koje u jeziku sa slobodnim redoslijedom riječi kao što je hrvatski ne možemo znati primjerice predstavlja li imenica direktni objekt (akuzativ), ili subjekt (nominativ).

(3) *Vani pada snijeg.*

(4) *Vani gledam snijeg.*

Primjer (3) daje sljedeća čitanja:

```
"<Vani>"
        "vani" adv
"<pada>"
        "padati" vblex iv prs p3 sg
"<snijeg>"
        "snijeg" n mi sg nom
        "snijeg" n mi sg acc
```

Vidimo da je glagol *pada* neprelazan (**iv**), tj. ne može imati objekt u akuzativu, pa riječ *snijeg* u (3) ne može biti njegov objekt. Stoga tu višeznačnost možemo razriješiti odabirom čitanja s nominativom. Pravilo kojim se to postiže je:

```
REMOVE Accusative IF (NOT 0* Verb + Transitive BARRIER BOS OR EOS)
```

Ovo pravilo uklanja čitanje s akuzativom u slučajevima kad u rečenici ne postoji prijelazni glagol. Ključne riječi BOS i EOS označavaju tokene za početak i kraj rečenice, a konstrukcija 0* označava skeniranje u oba smjera u odnosu na trenutni token. Ključna riječ BARRIER označava da se pretraga prekida ukoliko gramatika naiđe na njezin argument..

S druge strane, u (4):

```
"<Vani>"
        "vani" adv
"<gledam>"
        "gledati" vblex tv prs p1 sg
"<snijeg>"
        "snijeg" n mi sg nom
        "snijeg" n mi sg acc
```

glagol *gledam* je prijelazan i njegov direktni objekt je upravo *snijeg*, u akuzativu.

Slično, u rečenici (5) dolazi do višeznačnosti akuzativ/genitiv

(5) *Dozivam dobroga prijatelja.*

Čitanja za ovu rečenicu su:

```
"<Dozivam>"
        "dozivati" vblex tv prs p1 sg
"<dobrog>"
        "dobar" adj ma sg gen def
        "dobar" adj ma sg acc def
        "dobar" adj mi sg gen def
        "dobar" adj nt sg gen def
"<prijatelja>"
        "prijatelj" n ma sg gen
        "prijatelj" n ma sg acc
        "prijatelj" n ma pl gen
```

Kako je glagol *dozivam* prijelazan, imenska sintagma *dobroga prijatelja* treba dobiti čitanje u akuzativu, jer predstavlja objekt glagolu.

Pravilo za razrješavanje ovakve višeznačnosti je:

```
SELECT Accusative IF (0 Genitive OR Accusative) (-1 Verb + Transitive)
```

Gramatika prvo odabire čitanje u akuzativu za riječ *dobar*, jer se lijevo od nje nalazi prijelazni glagol. Zatim se niže opisanim pravilima za imenske sintagme ostvaruje potpuno morfološko razrješavanje.

Zanimljivo je napomenuti kako tranzitivnost i refleksivnost nisu morfološke kategorije, pa ova pravila ovise isključivo o glagolskim leksikonima CROVALLEX i apertium-sh-mk.

Imenske sintagme

Za imenske sintagme velik se broj višeznačnosti može lako razriješiti promatranjem slaganja u rodu, broju i padežu. To posebno vrijedi za prijedložne sintagme, koje zbog visokog stupnja višeznačnosti pridjevske deklinacije mogu imati vrlo velik broj čitanja. Svaki prijedlog u hrvatskom jeziku dolazi uz ograničen broj padeža i to se može iskoristiti da se smanji broj mogućih čitanja.

U primjeru (6), koji sadrži imensku sintagmu u lokativu, prijedlog *na* može se odnositi na dva padeža (lokativ/akuzativ), pridjevi *lijepom* i *plavom* imaju više čitanja za rod i tri padežna čitanja (dativ/lokativ/instrumental), a *Dunav* je višeznačan između dativa i lokativa.

(6) *Na lijepom plavom Dunavu.*

Čitanja za ovu rečenicu su dana u 2.2.
Ovdje se do razrješenja višeznačnosti dolazi primjenom niza pravila:

```
REMOVE $$Case IF (0 Preposition) (NOT 1 Nominal + $$Case)

REMOVE $$Case IF (0 Nominal)
                 (-1 Preposition + $$Case OR Modifier + $$Case)

SELECT $$Gender + $$Number
  IF (0 Modifier) (1 NP-HEAD + $$Gender + $$Number)
```

Prvo pravilo uklanja s prijedloga *"na"* sva čitanja osim onoga koje sadrži lokativ. Nakon toga drugo pravilo redom odabire lokativ kod svih preostalih riječi i eliminira čitanja s dativom. Treće pravilo prvom prolazu odabire rod i broj za pridjev *"plav"* s riječi *"Dunav"*, a u sljedećem prolazu odabire rod i broj za pridjev *"lijep"* od pridjeva *"plav"*.

Primjer 2.2: Primjer priložne sintagme

```
"<Na>"
        "na" pr loc
        "na" pr acc
"<lijepom>"
        "lijep" adj nt sg dat def
        "lijep" adj nt sg loc def
        "lijep" adj ma sg dat def
        "lijep" adj ma sg loc def
        "lijep" adj mi sg dat def
        "lijep" adj mi sg loc def
        "lijep" adj f sg ins ind
        "lijep" adj f sg ins def
"<plavom>"
        "plav" adj nt sg dat def
        "plav" adj nt sg loc def
        "plav" adj ma sg dat def
        "plav" adj ma sg loc def
        "plav" adj mi sg dat def
        "plav" adj mi sg loc def
        "plav" adj f sg ins ind
        "plav" adj f sg ins def
"<Dunavu>"
        "Dunav" n prp top mi sg dat
        "Dunav" n prp top mi sg loc
```

Simbol `$$` označava iteraciju po skupu. Primjerice, za `$$Number` pravilo se prvo izvršava za jedninu, a zatim za množinu. `Modifier` označava sve riječi koje se mogu slagati u rodu, broji ili padežu s nekom drugom imenskom riječi. `NP-HEAD` je kratica za glavu imenske sintagme.

Komplementarno prijedložnim sintagmama, možemo promotriti još jednu čestu višeznačnost - dativ/lokativ, koji su ortografski uvijek identični. Kako lokativ uvijek dolazi s prijedlogom, u slučajevima gdje sintagma nije prijedložna možemo jednostavno ukloniti čitanje s lokativom.

U sljedećoj sintagmi je očito riječ o dativu:

(7) *Plavom Dunavu!*

Dobivamo analize:

```
"<Plavom>"
        "plav" adj mi sg dat def
        "plav" adj mi sg loc def
"<Dunavu>"
        "Dunav" n prp top mi sg dat
        "Dunav" n prp top mi sg loc
```

Pravilo za razrješavanje je sljedeće:

```
REMOVE Locative IF (0 Nominal + Locative)
              (-1* Preposition + Locative BARRIER Word - Modifier)
```

Pravilo uklanja čitanja s lokativom ukoliko se lijevo od riječi unutar imenske sintagme ne nalazi prijedlog koji dolazi s lokativom. Barijera `Word - Modifier` označava riječi koje prekidaju imensku sintagmu.

Prepoznavanje vlastitih imenica

Jedno od jednostavnijih heurističkih pravila omogućuje prepoznavanje vlastitih imenica:

```
SUBSTITUTE $$Nominal $$Nominal + (np)
        TARGET ("<[A-Z].*>"r) + $$Nominal
        IF (0 Noun OR Adjective) (NOT -1 BOS)
```

Prefiks `$$` označava iteraciju po skupu svih imenskih riječi. Ova će heuristika dodijeliti oznaku vlastite imenice (**np**) pridjevu ili imenici koji se nalaze unutar rečenice i počinju velikim slovom. Pritom će se sačuvati i polazno čitanje kako bi ga se moglo iskoristiti u kasnijoj fazi.

Ovo pravilo otvara mogućnosti za dodatnu primjenu heuristika na kandidate za vlastitu imenicu. Primjerice, u sintagmi

(8) *...u glavnom gradu Lusaki ...*

s analizama:

```
"<u>"
        "u" pr acc
        "u" pr gen
        "u" pr loc
"<glavnom>"
        "glavni" adj f sg ins def
        "glavni" adj m sg dat def
        "glavni" adj m sg loc def
        "glavni" adj nt sg dat def
        "glavni" adj nt sg loc def
"<gradu>"
        "grad" n m sg dat
        "grad" n m sg loc
"<Lusaki>"
        "Lusaka" n f sg dat
        "Lusaka" n f sg loc
```

Temeljem prvog prijedloga lako možemo zaključiti da je cijela sintagma u lokativu. No, za riječ *Lusaka* s čitanjem opće imenice (**n**) ne možemo jednoznačno odrediti padež, jer se ispred nje nalazi druga imenica. Korištenjem gore navedene heuristike prvo možemo pogoditi da je riječ o vlastitoj imenici, a zatim na nju možemo primijeniti dodatnu heuristiku koja određuje da se ona slaže u broju i padežu s općom imenicom koja joj prethodi:

```
SELECT ProperNoun + $$Case + $$Number (-1 Noun + $$Case + $$Number)
```

Razrješavanje odnosnih zamjenica

Odnosna zamjenica slaže se u rodu i broju s imenicom na koju se odnosi. Primjerice u (9) zamjenica *kojima* odnosi se na riječ *hitovima*.

(9) *... **hitovima.m.pl** s prošla dva albuma, medju **kojima.m.pl** pjesmama ...*

Riječ *kojima* je višeznačna (može biti u ženskom, muškom ili srednjem rodu), dok je riječ *hitovima* jednoznačno imenica u množini muškog roda.

```
"<hitovima>"
        "hit" prn m pl ins
    ...
"<kojima>"
        "koji" prn f pl ins
        "koji" prn m pl ins
        "koji" prn nt pl ins
```

Odnosna zamjenica mora se odnositi na neku riječ slijeva, pa je možemo ispravno označiti heuristikama:

```
SELECT $$Gender IF (0 ("<kojima>"i) )
        (-1* Noun + Plural + $$Gender BARRIER ("koji"i))
SELECT $$Gender IF (0 ("<kojima>"i) )
        (-1* Pronoun + Plural + $$Gender BARRIER ("koji"i))
```

Oba pravila odabiru rod zamjenice *kojima*, tražeći nalijevo zamjenicu ili imenicu koja se s njom slaže u rodu i broju.

Zvjezdica i ključna riječ `BARRIER` označavaju da niz tokena pretražujemo nalijevo sve dok ne naiđemo na traženu riječ ili argument od `BARRIER`, koji je u ovom slučaju druga pokazna zamjenica *koji*.

Brojevne sintagme

Brojevne sintagme u hrvatskom jeziku dodatno kompliciraju čitanja, jer broj postaje glava sintagme, a ostatak sintagme dobiva različite padežne nastavke koji su ili ostatak stare deklinacije duala ili identični genitivu (i to genitivu jednine za brojeve koji završavaju na 2, 3, 4, a genitivu množine za ostale brojeve).

Prvi slučaj je morfološki složeniji, no razrješavanje je zbog toga jednostavnije. Ostatak deklinacije duala sadrži tri različita padežna nastavka. Nastavci za nominativ, akuzativ i vokativ identični su morfološkom genitivu (jednine ili množine). Dativ, lokativ i instrumental isti su kao u običnom pluralu, a genitiv dobiva poseban nastavak-*ju*.

(10) *Govorio je o dvjema ženama.*

Za ovaj slučaj višeznačnosti dovoljna su ranije pravila za prijedložne sintagme (prijedlog *o* dolazi s lokativom, pa cijela sintagma dobiva čitanje s lokativom).

Primjer drugog slučaja je rečenica (11). Broj se nalazi u morfološkom nominativu (zapravo se može odnositi na bilo koji padež), a ostatak sintagme je u genitivu. Ovaj je stil karakterističan za slobodnije tekstove, novinski ili razgovorni stil.

(11) *Govorio je o dvije žene.*

U ovom slučaju se članovima brojevne sintagme (s izuzetkom broja) dodjeljuje genitiv jednine ili množine, ovisno o broju, a sam broj i prijedlog se razrješavaju neovisno o njima.

Oba slučaja su različito pokrivena morfološkim analizatorom. Broj dobiva pune padežne oznake ili mu se jednostavno dodjeljuje čitanje bez padeža. Kako nema općenitog načina za određivanje koji od ova dva stila se koristi u tekstu, nužno je uzeti u obzir sve moguće kombinacije.

```
REMOVE $$Case IF (0 Num)
   (-1* Prep BARRIER WORD - Modifier)
   (NOT -1* Prep + $$Case BARRIER WORD - Modifier)

SELECT Genitive IF (0 Nominal)
              (0* Numeral + Nominative BARRIER WORD - Modifier)
```

Prvo pravilo iterira kroz skup padeža i za svaki uklanja s glave brojevne sintagme ukoliko joj ne prethodi prijedlog koji dolazi s tim padežom. Ovo pravilo u kombinaciji s pravilima za prijelazne glagole i akuzativ uklanja sve višeznačnosti koje je moguće razriješiti na nesemantički način. Drugo pravilo odabire čitanja s genitivom ukoliko se unutar sintagme nalazi broj u nominativu.

3 Prijedlog proširenja za analizu na sintaktičkoj razini

3.1 Sintaktička analiza u gramatici ograničenja

U klasičnoj obradi teksta gramatikom ograničenja korak koji slijedi nakon razrješavanja morfoloških višeznačnosti jest sintaktička analiza. Sintaktičko označavanje nije moguće ispravno izvesti bez potpune razriješenosti svih morfosintaktičkih višeznačnosti, a i postupak je sam po sebi otprilike jednake složenosti, pa daleko prelazi opseg ovog rada. Stoga ćemo ovdje samo ukratko opisati kako taj postupak funkcionira.

Sintaktička analiza u gramatici ograničenja podrazumijeva sličan slijed koraka kao kod razrješavanja morfoloških višeznačnosti. Obično se odvija u dvije faze:

- dodjeljivanje sintaktičkih oznaka,
- razrješavanje sintaktičkih višeznačnosti.

U prvoj se fazi pojavnicama dodjeljuju sve moguće sintaktičke oznake, ovisno o njihovoj jednoznačnoj oznaci vrste riječi ili morfološkoj kategoriji.

U drugom se koraku sasvim analogno fazi razrješavanja morfoloških višeznačnosti na kohorte primjenjuje niz pravila koja razrješavaju sintaktičke višeznačnosti.

Sintaktička analiza gramatike ograničenja temelji se na skupu funkcijskih oznaka, koje označavaju sintaktičku funkciju pojedine riječi, a mogu sadržavati i pokazivač ovisnosti - '<' ili '>', koji pokazuje ovisi li riječ o nekoj riječi slijeva ili zdesna. One također mogu biti proizvoljne, no počinju *funkcijskim prefiksom* (standardno `@`). Primjer nekih standardnih funkcijskih oznaka za sintaktičku analizu može se naći u tablici 3.1.

Oznake sintaktičkih oznaka dodaju se riječima uporabom pravila oblika `MAP`:

```
MAP (@SUBJ< @SUBJ>) TARGET Noun + Nominative
```

Ovo pravilo svakoj imenici u nominativu dodaje oznaku subjekta (oznake smjera se odnose na položaj predikata u rečenici.)

Tablica 3.1: Standardne sintaktičke oznake gramatike ograničenja

Oznaka	Naziv	Oznaka	Naziv
@SUBJ>	Subjekt predikatu zdesna	@CO	koordinator
@SUBJ<	Subjekt predikatu slijeva	@SUB	subordinator
@ACC	Direktni objekt (akuzativ)	@TOP	tema
@DAT	Indirektni objekt (dativ)	@APP	apozicija
@PIV	Prijedložni objekt	@>N	ovisi o imenskoj riječi zdesna
@SC	Komplement subjektu	@<N	ovisi o imenskoj riječi slijeva
@OC	Komplement objektu	@N<PRED	predikat imenskoj riječi slijeva
@SA	arg. prijedloga koji se odnosi na subjekt	@>A	ovisi o prilogu zdesna
@OA	arg. prijedloga koji se odnosi na objekt	@<A	ovisi o prilogu slijeva
@MV	glavni glagol	@P<	argument prijedloga slijeva
@AUX	pomoćni glagol	@>>P	argument zalutalog prijedloga zdesna
@ADVL	priložna oznaka	@INFM	infinitivni marker
@AUX<*	argument pomoćnog glagola	@VOC	vokativ
@PRED	predikat	@FOC	fokalni marker

*oznake < i > znače da je trenutna riječ u odnosu s nekom riječi slijeva ili zdesna

Funkcijske oznake razlikuju se od običnih po tome što gramatika ograničenja dopušta da svako čitanje sadrži najviše jednu funkcijsku oznaku. Stoga, ako jednom čitanju dodijelimo više funkcijskih oznaka, gramatika će ih razdvojiti u više čitanja.

3.2 Jednostavan primjer razrješavanja

Sintaktičke višeznačnosti možemo ilustrirati na jednostavnom primjeru (3.1):

Posvojna zamjenica *Naša* i pridjev *debela* dobili su oznake `@N>` i `@N<`, jer se moraju odnositi na neku imensku riječ. Riječ *mačka* je u nominativu, pa je dobila oznake subjekta `@SUBJ>` i `@SUBJ<`. Riječ *stalno* je prilog, pa smo je proglasili "priložnom oznakom" (`@ADVL>`/`@ADVL<`). Na koncu, glagol *jesti* i imenica *riba* su prijelazan glagol odnosno njegov bliži (izravni objekt) pa su dobili oznake glavnog glagola (`@FMV>`/`@FMV<`) odnosno oznaku direktnog objekta (`@ACC>`/`@ACC<`).

Za razrješavanje dobivenih višeznačnosti mogli bismo napisati nekoliko jednostavnih pravila 3.2.[1]

Prvo pravilo za riječ iz klase `Modifier` (*naš* i *debeo*) odabire `@N>` jer glava imenske sintagme s kojom se slažu u rodu, broju i padežu nalazi zdesna. `BARRIER` sadrži riječi koje prekidaju imensku sintagmu. Drugo pravilo odabire `@SUBJ>` za imenicu *mačka* jer se njezin predikat nalazi njoj zdesna (glagol koji se s njom slaže u broju). Treće pravilo za prilog

[1] Pravila služe samo za jednostavnu ilustraciju, i svakako ne pokrivaju kompliciranije slučajeve

Primjer 3.1: Primjer za rečenicu *Naša debela mačka stalno jede ribu*

```
"<Naša>"
    "naš" prn pos f sg nom @N>
    "naš" prn pos f sg nom @N<
"<debela>"
    "debeo" adj f sg nom @N>
    "debeo" adj f sg nom @N<
"<mačka>"
    "mačka" n f sg nom @SUBJ>
    "mačka" n f sg nom @SUBJ<
"<stalno>"
    "stalno" adv @ADVL>
    "stalno" adv @ADVL<
"<jede>"
    "jesti" vblex imperf tv p3 sg @FMV>
    "jesti" vblex imperf tv p3 sg @FMV<
"<ribu>"
    "riba" n f sg acc @<ACC
    "riba" n f sg acc @>ACC
```

Primjer 3.2: Primjer pravila za razrješavanje sintaktičkih višeznačnosti

```
SELECT Modifier + (@N>) IF
    (0 Modifier + $$Gender + $$Number + $$Case)
    (1* Noun + $$Gender + $$Number + $$Case
    BARRIER Word - Modifier)
SELECT Noun + $$Number + (@SUBJ>) IF
    (1* Verb + $$Number BARRIER Verb - $$Number)
SELECT Adverb + (@ADVL>) IF (1 Verb)
SELECT Verb + (@FMV>) IF (1 Noun + Accusative)
SELECT Noun + (@ACC<) IF (-1 Verb + Transitive)
```

Primjer 3.3: Rečenica s razriješenim sintaktičkim višeznačnostima

```
"<Naša>"
    "naš" prn pos f sg nom @N>
"<debela>"
    "debeo" adj f sg nom @N>
"<mačka>"
    "mačka" n f sg nom @SUBJ>
"<stalno>"
    "stalno" adv @ADVL>
"<jede>"
    "jesti" vblex imperf tv p3 sg @FMV>
"<ribu>"
    "riba" n f sg acc @<ACC
```

stalno odabire @ADVL> jer se glagol na koji se odnosi nalazi njemu zdesna. Posljednja dva pravila iskorištavaju odnos prijelazni glagol - objekt u akuzativu i odabire oznake @FMV> i @ACC< za glagol *jesti* i njegov direktni objekt *riba*. Rečenica s potpuno razriješenim sintaktičkim čitanjima prikazana je u primjeru 3.3

Kao što je vidljivo, primjerice, iz oznaka za direktni i indirektni objekt (@ACC, @DAT), koji se u hrvatskom izražavaju direktno morfološki kroz dativ i akuzativ, ovaj skup oznaka razvijan je prvenstveno za jezike u kojima su to sintaktičke, a ne morfološke kategorije. Za ozbiljnije sintaktičko označavanje hrvatskih tekstova svakako bi trebalo prilagoditi skup oznaka hrvatskoj jezičnoj tradiciji.

4 Vrednovanje

4.1 Opis označenog korpusa

Trenutna implementacija gramatike ograničenja vrednovana je temeljem usporedbe s ručno označenim skupom rečenica (tzv. zlatnim standardom) od otprilike 5000 pojavnica. Skup je dobiven nasumičnim odabirom rečenica iz korpusa novinskih članaka *Vjesnika*, odnosno istog korpusa koji je korišten za razvoj gramatike, i nije korišten za dodatno poboljšavanje pravila gramatike.

Kako je neke višeznačnosti bilo nemoguće razriješiti na morfološkoj razini (semantičke višeznačnosti) ostavljene su nerazriješene i u zlatnom standardu. Najčešći primjeri toga su dativ/instrumental množine

```
"<sindikalistima>"
        "sindikalist" n m pl dat
        "sindikalist" n m pl ins
...
"<rukama>"
        "ruka" n f pl dat
        "ruka" n f pl ins
```

i genitiv jednine/množine

```
"<radnika>"
        "radnik" n m sg gen
        "radnik" n m pl gen
...
"<tvornica>"
        "tvornica" n f sg gen
        "tvornica" n f pl gen
```

4.2 Metode vrednovanja

Performanse sustava vrednovane su korištenjem uobičajenih mjera iz pretraživanja informacija – preciznosti (P), odaziva (R), i F-mjere. Preciznost označava postotak čitanja u izlazu gramatike koja su točna, a odaziv postotak čitanja iz zlatnog standarda koja se nalaze u izlazu gramatike. F-mjera je njihova harmonijska sredina.

$$P = \frac{|\text{gold} \cap \text{cg}|}{|\text{cg}|}$$

$$R = \frac{|\text{gold} \cap \text{cg}|}{|\text{gold}|}$$

$$F = 2 \times \frac{P \cdot R}{P + R}$$

Ovdje *cg* označava skup čitanja nakon obrade gramatikom, a *gold* označava skup čitanja koja sadrži zlatni standard. Presjeke skupova smo računali na razini čitanja. Uzimamo da su dva čitanja jednaka ako se poklapaju u lemi i svim morfosintaktičkim oznakama. Kao dodatna mjera korišten je postotak razriješenosti čitanja.

Sustav smo vrednovali na dva načina: označavanje vrste riječi (13 oznaka) i označavanje uporabom potpunog skupa morfosintaktičkih oznaka (cijeli skup iz 2.1).

4.3 Rezultati

Rezultati vrednovanja dani su u tablici 4.1. Čitanja koja su namjerno ostavljena nerazriješenima u zlatnom standardu dovode kod riječi koje označivač nije posve razriješio do veće preciznosti u odnosu na idealnu preciznost.

Tablica 4.1: Performanse označivača (%)

	P	R	F-score	Razriješeno
POS tagging	96.08	99.76	97.88	95.30
Morphosyntactic tagging	88.16	98.13	92.88	86.36

4.4 Analiza pogrešaka

Iako su rezultati dobiveni za relativno malen broj pravila obećavajući, usporedba s rezultatima razvijenijih gramatika za druge jezike pokazuje kako se na gramatici svakako još može puno poraditi. Glavni problem predstavlja nerazriješenost znatnog broja čitanja, koja dovodi do relativno niske preciznosti. To se može pripisati nekoliko uzroka. Prvenstveno, gramatika je još u razvoju. Neka pravila nisu temeljito ispitana, a pojedine gramatičke konstrukcije nisu u potpunosti pokrivene. Primjer toga je vrlo česta višeznačnost nominativ/akuzativ, kojom se pravila tek djelomično bave. Analize s akuzativom uklanjaju se ako u rečenici ne postoji prijelazni glagol, a odnos glagol/objekt uzima se u obzir samo ako se glagol nalazi odmah slijeva imenici. Trenutno ne postoje nikakva pravila koja bi uzela u obzir udaljenije relacije subjekt/objekt u odnosu na nominativ/akuzativ. Također, spomenuta pravila ovise o glagolskim leksikonima CROVALLEX i `apertium-sh-mk`. Slučajevi kad se pojavi glagol čija valencija je nepoznata uopće nisu pokriveni i predstavljaju vrlo težak problem zbog slobodnog reda riječi u hrvatskom jeziku (primjer 4.1).

Drugi problem predstavljaju morfološke analize. Morfološki je leksikon dobiven poluautomatskom akvizicijom iz korpusa. Iako ima vrlo veliku pokrivenost, nekim riječima su pridružene pogrešne analize, dok su neke analize nepotpune (primjer 4.2). Netočna ili nepotpuna analiza pojedine riječi ometa ispravno razrješavanje te riječi, no često također narušava razrješavanje riječi u njezinoj neposrednoj okolini.

Na koncu, znatan broj nepotpuno razriješenih čitanja otpada na nepoznate riječi4.3, budući da naš analizator trenutno ne podržava njihovu analizu. Nepoznate riječi posebno su nezgodne kad se pojave unutar imenskih sintagmi, gdje prekidaju tok rada gramatike.

Primjer 4.1: Primjer glagola s nepoznatom valencijom (ugovoriti, tv?)

```
"<Ugovorimo>"
        "ugovoriti" vblex imp p1 pl tv? SELECT:955:isImperative
;        "ugovoriti" vblex pres p1 pl tv? SELECT:955:isImperative
"<sastanak>"
        "sastanak" n m sg acc
        "sastanak" n m sg nom
"<s>"
        "s" pr ins
;        "s" pr gen REMOVE:497:PrPhrase_Preposition_Case_Cleaning
"<hrvatskim>"
        "hrvatski" adj pos m sg ins def @N→ ADD:361:@N→
;        "hrvatski" adj pos f pl dat def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos f pl loc def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos m pl dat def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos m pl loc def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos nt pl dat def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos nt pl loc def REMOVE:504:PrPhrase_Constituent_Case_Cleaning_Direct
;        "hrvatski" adj pos nt pl ins def REMOVE:543:Modifier_Cleaning_fromNoun_Direct
;        "hrvatski" adj pos f pl ins def REMOVE:543:Modifier_Cleaning_fromNoun_Direct
;        "hrvatski" adj pos nt sg ins def REMOVE:543:Modifier_Cleaning_fromNoun_Direct
;        "hrvatski" adj pos m pl ins def REMOVE:543:Modifier_Cleaning_fromNoun_Direct
"<jezikom>"
        "jezik" n m sg ins
```

Primjer 4.2: Primjer pogrešne analize (tvrtka, nedostaje množina)

```
"<14>"
        "14" num crd @NumPhC %Plural ADD:325:ItsANumber ADD:327:ItsANumber
"<domaćih>"
        "domaći" adj f pl gen def
        "domaći" adj m pl gen def
        "domaći" adj nt pl gen def
"<tvrtki>"
        "tvrtka" n f sg dat
;        "tvrtka" n f sg loc REMOVE:642:NoPreposition
"<.>"
        "." fxq
```

Primjer 4.3: Primjer nepoznate riječi (lokativ ostaje nerazriješen)

```
"<Na>"
        "na" pr loc
        "na" pr acc
"<samom>"
        <unk>
"<sjeveru>"
        "sjever" n m sg loc
        "sjever" n m sg dat
"<Njemačke>"
        "Njemačka" np pos f sg gen def SUBSTITUTE:283:[H][Mod]ProperNoun SUBSTITUTE:839:[H][Mod]NpAdj_isNP
;        "Njemačka" adj np pos f pl voc def SUBSTITUTE:283:[H][Mod]ProperNoun REMOVE:633:NotVocative
;        "Njemačka" np pos f pl acc def SUBSTITUTE:283:[H][Mod]ProperNoun SUBSTITUTE:839:[H][Mod]NpAdj_isNP
;        "Njemačka" np pos f pl nom def SUBSTITUTE:283:[H][Mod]ProperNoun SUBSTITUTE:839:[H][Mod]NpAdj_isNP
;        "Njemačka" np pos m pl acc def SUBSTITUTE:283:[H][Mod]ProperNoun SUBSTITUTE:839:[H][Mod]NpAdj_isNP
```

5 Zaključak

U ovome radu opisali smo problematiku označavanja riječi i razrješavanja morfosintaktičke višeznačnosti u kontekstu obrade prirodnog jezika. Nabrojali smo neke pristupe rješavanju tog problema te opisali gramatiku ograničenja kao učinkovitu paradigmu za razrješavanje morfosintaktičke višeznačnosti zasnovanu na ručno pisanim pravilima. Razmotrili i smo problem označavanja na razini sintaktičkih funkcija.

Opisan je razvijeni prototip morfosintaktičkog označivača za hrvatski jezik zasnovanog na gramatici ograničenja. Označivač koristi opširan skup morfosintaktičkih oznaka (kako i priliči jeziku bogate morfologije kao što je hrvatski) i morfološki analizator temeljen na flektivnom leksikonu u kombinaciji s dvama leksikonima glagolskih valencija.

Naša gramatika trenutno sadrži 290 pravila, što je malo u usporedbi sa zrelijim sustavima temeljenima na gramatici ograničenja. Prema očekivanjima, iz rezultata vrednovanja se da zaključiti kako će za zadovoljavajuće označavanje biti potreban daljnji i detaljniji razvoj, uz precizniju morfološku analizu.

Međutim, s obzirom na to da je ovo je prvi rad u kojem je ovakva paradigma ozbiljno primjenjena na neki slavenski jezik (ne računajući gramatike za ruski i makedonski jezik u okviru Apertiuma, koje su vrlo male i nisu ih razvijali izvorni govornici) rezultati su ohrabrujući. Uz relativno malo pravila postignuta je prilično visoka preciznost označavanja vrste riječi i zadovoljavajuća preciznost pri punom morfosintaktičkom označavanju.

Kako se u hrvatskom jeziku često pojavljuju višeznačnosti koje je nemoguće razriješiti na morfosintaktičkoj razini, već je za njihovo razrješavanje potrebno imati dodatno znanje, na primjer semantičko znanje ili podatke o glagolskim valencijama, uz trenutno korišteni skup oznaka sve višeznačnosti nije moguće razriješiti samo ručno pisanim pravilima. No, iz brzine kojom se broj višeznačnosti smanjivao s uvođenjem novih pravila intuitivno se da zaključiti kako bi se uz relativno malo truda mogla razviti baza pravila koju bi bilo lako kombinirati s nekim stohastičkim označivačem, te da bi takva kombinacija mogla dati vrhunske rezultate.

Valja napomenuti i kako je gramatika razvijena na korpusu novinskog tipa i ispitana na njegovom vrlo malom isječku, koji je za ovu priliku morao biti ručno označen. Detaljnijem vrednovanju, a i više razgranatom razvoju gramatike, svakako bi pogodovao veći označeni korpus iz različitih domena na kojem bi bilo moguće razrađivati elaborantnija pravila.

6 Dodaci

6.1 Dodatak A

Djelovanje pravila za akuzativ i prijelazne/neprelazne glagole

```
"<imaju>"
    "imati" vblex pres p3 pl tv imperf
;   "imati" vblex pres p3 pl imperf iv REMOVE:VerbWithATransitiveObject
"<zabavnu>"
    "zabavan" adj f sg acc def? @N→ SUBSTITUTE:[H][Mod]definite/indefinite? ADD:@N→
;   "zabavan" adj f sg acc ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "zabavan" adj nt sg loc ¢IND ADD:[Mod]definite/indefinite? REMOVE:Modifier_Cleaning_fromNoun
;   "zabavan" adj m sg dat ¢IND ADD:[Mod]definite/indefinite? REMOVE:Modifier_Cleaning_fromNoun
;   "zabavan" adj m sg loc ¢IND ADD:[Mod]definite/indefinite? REMOVE:Modifier_Cleaning_fromNoun
;   "zabavan" adj nt sg dat ¢IND ADD:[Mod]definite/indefinite? REMOVE:Modifier_Cleaning_fromNoun
"<crtu>"
    "crta" n f sg acc
...
"<reagirao>"
    "reagirati" vblex lp m sg dual iv SUBSTITUTE:[Mod]LParticiple
"<kordon>"
    "kordon" n m sg nom
;   "kordon" n m sg acc REMOVE:[H]NoTransitiveVerb
"<vojnih>"
    "vojni" adj m pl gen def @N→ ADD:@N→
;   "vojni" adj f pl gen def REMOVE:Modifier_Cleaning_fromNoun
;   "vojni" adj nt pl gen def REMOVE:Modifier_Cleaning_fromNoun
"<policajaca>"
    "policajac" n m pl gen
;   "policajac" n m sg gen REMOVE:Noun_Cleaning_byModifier
...
"<To>"
    "to" prn dem nt sg nom adj
;   "to" prn dem nt sg acc adj REMOVE:[H]NoTransitiveVerb
"<su>"
    "biti" vbcop pres p3 pl
"<naklapanja>"
    "naklapanje" n nt pl nom
;   "naklapanje" n nt pl gen REMOVE:Noun_Cleaning_byModifier
;   "naklapanje" n nt sg gen REMOVE:Noun_Cleaning_byModifier
;   "naklapanje" n nt pl voc REMOVE:NotVocative
;   "naklapanje" n nt pl acc REMOVE:[H]NoTransitiveVerb
```

Djelovanje pravila za višeznačnost prilog/pridjev:

```
"<Bilo>"
    "biti" vbcop sg nt lp SELECT:[H]BiBilo
;   "biti" vblex lp nt sg tv imperf dual SUBSTITUTE:[Mod]LParticiple SELECT:[H]BiBilo
;   "biti" vblex lp nt sg imperf dual iv SUBSTITUTE:[Mod]LParticiple SELECT:[H]BiBilo
"<bi>"
    "biti" vbcop aor p3 sg SELECT:ConditionalNumber SELECT:[H]BiloBiX
;   "biti" vbcop aor p3 pl SELECT:ConditionalNumber
;   "biti" vbcop aor p2 sg SELECT:ConditionalNumber SELECT:[H]BiloBiX
"<dobro>"
    "dobro" adv
;   "dobar" adj nt sg acc ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "dobar" adj nt sg nom ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "dobar" adj nt sg voc ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "dobar" adj nt sg voc def? @Mod→ SUBSTITUTE:[H][Mod]definite/indefinite? ADD:Mod→Mod REMOVE:NotVocative
;   "dobar" adj nt sg acc def? @Mod→ SUBSTITUTE:[H][Mod]definite/indefinite? ADD:Mod→Mod REMOVE:[H]Adv|Adj
;   "dobar" adj nt sg nom def? @Mod→ SUBSTITUTE:[H][Mod]definite/indefinite? ADD:Mod→Mod REMOVE:[H]Adv|Adj
"<javno>"
    "javno" adv
;   "javan" adj nt sg acc ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "javan" adj nt sg nom ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "javan" adj nt sg voc ¢IND ADD:[Mod]definite/indefinite? REMOVE:[H][Mod]definite/indefinite?
;   "javan" adj nt sg voc def? SUBSTITUTE:[H][Mod]definite/indefinite? REMOVE:NotVocative
;   "javan" adj nt sg acc def? SUBSTITUTE:[H][Mod]definite/indefinite? REMOVE:[H]Adv|Adj
;   "javan" adj nt sg nom def? SUBSTITUTE:[H][Mod]definite/indefinite? REMOVE:[H]Adv|Adj
"<upitati>"
    "upitati" vblex inf perf iv
;   "upitati" vblex inf ref perf REMOVE:no_Reflexive_Pronoun
"<takve>"
    "takav" prn dem m pl acc adj @N→ SUBSTITUTE:[Mod]ImpliedMasculine ADD:@N→
;   "takav" prn dem f pl acc adj REMOVE:Modifier_Cleaning_fromNoun
;   "takav" prn dem f pl nom adj REMOVE:Modifier_Cleaning_fromNoun
;   "takav" prn dem f pl voc adj REMOVE:Modifier_Cleaning_fromNoun
;   "takav" prn dem f sg gen adj REMOVE:Modifier_Cleaning_fromNoun
"<dužnosnike>"
    "dužnosnik" n m pl acc
"<i>"
    "i" cnj coo
"<»>"
    "»" lqt
"<stručnjake>"
    "stručnjak" n m pl acc
"<«>"
    "«" rqt
```

Djelovanje pravila za prijedložne sintagme

```
"<u>"
    "u" pr acc
    "u" pr loc
;   "u" pr gen REMOVE:PrPhrase_Preposition_Case_Cleaning
"<tim>"
    "taj" prn dem f pl loc adj @N→ ADD:@N→ SUBSTITUTE:[Mod]ThisIsLocative SUBSTITUTE:[Mod]ThisIsLocative
    "tim" n m sg acc
;   "taj" prn dem f pl dat adj @N→ ADD:@N→ REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem f pl ins adj @N→ ADD:@N→ REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem m pl dat adj SUBSTITUTE:[Mod]ImpliedMasculine REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem m pl ins adj SUBSTITUTE:[Mod]ImpliedMasculine REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem m sg ins adj SUBSTITUTE:[Mod]ImpliedMasculine REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem nt pl dat adj REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem nt pl ins adj REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem nt sg ins adj REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "tim" n m sg nom REMOVE:PrPhrase_Constituent_Case_Cleaning
;   "taj" prn dem nt pl loc adj REMOVE:Modifier_Cleaning_fromNoun
;   "taj" prn dem m pl loc adj SUBSTITUTE:[Mod]ImpliedMasculine REMOVE:Modifier_Cleaning_fromNoun
"<županijama>"
    "županija" n f pl loc SUBSTITUTE:[Mod]ThisIsLocative SUBSTITUTE:[Mod]ThisIsLocative
;   "županija" n f pl dat REMOVE:Noun_Cleaning_byModifier
;   "županija" n f pl ins REMOVE:Noun_Cleaning_byModifier
...
"<Na>"
    "na" pr loc
;   "na" pr acc REMOVE:PrPhrase_Preposition_Case_Cleaning
"<samom>"
    "unk" <unk>
"<sjeveru>"
    "sjever" n m sg loc
;   "sjever" n m sg dat REMOVE:PrPhrase_Constituent_Case_Cleaning
"<Njemačke>"
    "Njemačka" np pos f sg gen def SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]NpAdj
;   "Njemačka" adj np pos f pl voc def SUBSTITUTE:[H][Mod]ProperNoun REMOVE:NotVocative
;   "Njemačka" np pos f pl acc def SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]NpAdj REMOVE:[H]NP_SG
;   "Njemačka" np pos f pl nom def SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]NpAdj REMOVE:[H]NP_SG
;   "Njemačka" np pos m pl acc def SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]NpAdj REMOVE:[H]NP_SG
```

Djelovanje pravila za vlastite imenice

```
"<zagrebačkog>"
    "zagrebački" adj pos m sg gen def @N→ ADD:@N→ ADD:Mod→Mod
    "zagrebački" adj pos nt sg gen def @N→ ADD:@N→ ADD:Mod→Mod
;   "zagrebački" adj pos m sg gen def @Mod→ ADD:@N→ ADD:Mod→Mod REMOVE:Modifier_Cleaning_fromNoun
;   "zagrebački" adj pos nt sg gen def @Mod→ ADD:@N→ ADD:Mod→Mod REMOVE:Modifier_Cleaning_fromNoun
"<Županijskog>"
    "Županijski" adj np pos m sg gen def @N→ SUBSTITUTE:[H][Mod]ProperNoun ADD:@N→
;   "Županijski" adj np pos nt sg gen def SUBSTITUTE:[H][Mod]ProperNoun REMOVE:Modifier_Cleaning_fromNoun
"<suda>"
    "sud" n m sg gen
...
"<odredilo>"
    "odrediti" vblex lp nt sg tv perf SUBSTITUTE:[Mod]LParticiple
"<pritvor>"
    "pritvor" n m sg acc
    "pritvor" n m sg nom
"<Stjepanu>"
    "Stjepan" np m sg dat SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]ProperNounCleanup
;   "Stjepan" np m sg loc SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]ProperNounCleanup REMOVE:NoPreposition
"<Mihaljincu>"
    "Mihaljinac" np m sg dat SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]ProperNounCleanup
;   "Mihaljinac" np m sg loc SUBSTITUTE:[H][Mod]ProperNoun SUBSTITUTE:[H][Mod]ProperNounCleanup REMOVE:NoPreposition
```

Djelovanje pravila za odnosne zamjenice

```
"<album>"
    "album" n m sg acc
    "album" n m sg nom
"<koji>"
    "koji" prn ind m sg acc adj SUBSTITUTE:[Mod]MiMa→M SELECT:koji
    "koji" prn ind m sg nom adj SELECT:koji
;   "koji" prn ind m pl nom adj SELECT:koji
"<će>"
    "htjeti" vbaux clt pres p3 pl SUBSTITUTE:[Mod]VBAux_Cleanup SUBSTITUTE:[Mod]VBAux_Cleanup SUBSTITUTE:[Mod]VBAux_Cleanup
    "htjeti" vbaux clt pres p3 sg SUBSTITUTE:[Mod]VBAux_Cleanup SUBSTITUTE:[Mod]VBAux_Cleanup SUBSTITUTE:[Mod]VBAux_Cleanup
;   "htjeti" vbaux pres p3 pl imperf iv REMOVE:[Mod]VBAux_Cleanup
;   "htjeti" vbaux pres p3 sg imperf iv REMOVE:[Mod]VBAux_Cleanup
"<najaviti>"
    "najaviti" vblex inf tv perf
;   "najaviti" vblex inf ref perf REMOVE:no_Reflexive_Pronoun
"<single>"
    "single" n m pl acc
    "single" n m sg nom
    "single" n m sg acc
;   "single" n m sg voc REMOVE:NotVocative
```

Djelovanje pravila za brojevne sintagme

```
"<hitova>"
    "hit" n m pl gen
"<s>"
    "s" pr gen @NumPhPr %Dual ADD:LongDist ADD:DualNumPh
    "s" pr ins @NumPhPr %Dual ADD:LongDist ADD:DualNumPh
"<prošla>"
    "prošli" adj m sg gen ¢IND %Dual @Mod→ ADD:[Mod]def/indef? ADD:Direct_FromPr ADD:Mod→Num
    "prošli" adj nt sg gen ¢IND %Dual @Mod→ ADD:[Mod]def/indef? ADD:Direct_FromPr ADD:Mod→Num
;   "prošli" adj f sg nom ¢IND ADD:[Mod]def/indef? REMOVE:[H][Mod]def/indef?
;   "prošli" adj f sg voc ¢IND ADD:[Mod]def/indef? REMOVE:[H][Mod]def/indef?
;   "prošli" adj nt pl acc ¢IND ADD:[Mod]def/indef? REMOVE:[H][Mod]def/indef?
;   "prošli" adj nt pl nom ¢IND ADD:[Mod]def/indef? REMOVE:[H][Mod]def/indef?
;   "prošli" adj nt pl voc ¢IND ADD:[Mod]def/indef? REMOVE:[H][Mod]def/indef?
;   "proći" vblex lp f sg perf iv SUBSTITUTE:[Mod]LParticiple REMOVE:NominalAfterPr
;   "proći" vblex lp f sg tv perf SUBSTITUTE:[Mod]LParticiple REMOVE:NominalAfterPr
;   "prošli" adj f sg nom def? @NumPhC %Dual SUBSTITUTE:[H][Mod]def/indef? ADD:Direct_FromPr
;   "prošli" adj nt pl voc def? @NumPhC %Dual SUBSTITUTE:[H][Mod]def/indef? ADD:Direct_FromPr
;   "prošli" adj f sg voc def? @NumPhC %Dual SUBSTITUTE:[H][Mod]def/indef? ADD:Direct_FromPr
;   "prošli" adj nt pl acc def? @NumPhC %Dual SUBSTITUTE:[H][Mod]def/indef? ADD:Direct_FromPr
;   "prošli" adj nt pl nom def? @NumPhC %Dual SUBSTITUTE:[H][Mod]def/indef? ADD:Direct_FromPr
;   "prošli" adj m sg gen ¢IND %Dual @NumPhC ADD:[Mod]def/indef? ADD:Direct_FromPr ADD:Mod→Num
;   "prošli" adj nt sg gen ¢IND %Dual @NumPhC ADD:[Mod]def/indef? ADD:Direct_FromPr ADD:Mod→Num
"<dva>"
    "dva" num crd pl ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber
;   "dva" num crd m pl acc ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd m pl nom ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd m pl voc ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd nt pl acc ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd nt pl nom ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd nt pl voc ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum_LongDist
;   "dva" num crd nt pl ltr @NumPhC %Dual ADD:ItsANumber ADD:ItsANumber REMOVE:ØCaseNum
"<albuma>"
    "album" n m sg gen @NumPhC %Dual ADD:Direct_FromNumLeft ADD:Direct_FromNumLeft SELECT:DualNum[ltr]_Left_Direct
;   "album" n m pl gen @NumPhC %Dual ADD:Direct_FromNumLeft ADD:Direct_FromNumLeft SELECT:DualNum[ltr]_Left_Direct
```

6.2 Dodatak B

Isječak iz ručno označenog korpusa

```
"<Delegati>"
    "delegat" n m pl nom
"<više>"
    "visoko" adv comp
"<od>"
    "od" pr gen
"<40>"
    "40" num crd
"<afričkih>"
    "afrički" adj pos f pl gen def
"<država>"
    "država" n f pl gen
"<okupljenih>"
    "okupljen" adj f pl gen def?
"<na>"
    "na" pr loc
"<sastanku>"
    "sastanak" n m sg loc
"<u>"
    "u" pr loc
"<zambijskome>"
    "zambijski" adj m sg loc def
"<glavnom>"
    "glavni" adj m sg loc def
"<gradu>"
    "grad" n m sg loc
"<Lusaki>"
    "Lusaka" np f sg loc
"<izabrali>"
    "izabrati" vblex lp m pl tv perf
"<su>"
    "biti" vbcop pres p3 pl
"<bivšega>"
    "bivši" adj m sg acc def
"<ministra>"
    "ministar" n m sg acc
"<vanjskih>"
    "vanjski" adj pos m pl gen def
"<poslova>"
    "posao" n m pl gen
"<Amaru>"
    "Amar" np m sg acc
"<Essyja>"
    "Essy" np m sg acc
"<za>"
    "za" pr acc
"<prijelaznoga>"
    "prijelazni" adj m sg acc def
"<glavnog>"
    "glavan" adj m sg acc def
"<tajnika>"
    "tajnik" n m sg acc
"<na>"
    "na" pr acc
"<mjesto>"
    "mjesto" n nt sg acc
"<dosadašnjega>"
    "dosadašnji" adj m sg gen def
"<Salima>"
    "Salim" np m sg gen
"<Ahmeda>"
    "Ahmed" np m sg gen
"<Salima>"
    "Salim" np m sg gen
"<.>"
    "." fxq sent
```

[]

Bibliografija

[1] Aduriz, I., J. M. Arriola, X. Artola, A. D. de Ilarraza *et al.*: *Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism*. U *Proceedings of RANLP97*, stranice 282–288, 1997.

[2] Agić, Ž. i M. Tadić: *Evaluating Morphosyntactic Tagging of Croatian Texts*. U *Proc. of the 5th Int. Conference on Language Resources and Evaluation*, 2006.

[3] Agić, Ž., M. Tadić i Z. Dovedan: *Improving part-of-speech tagging accuracy for Croatian by morphological analysis*. Informatica, 32(4):445–451, 2008.

[4] Bick, E.: *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus Univ. Press, 2000.

[5] Bick, E.: *A CG & PSG Hybrid Approach to Automatic Corpus Annotation*. U *Proceedings of SProLaC2003*, stranice 1–12, 2003.

[6] Bick, E.: *Parsing and Evaluating the French Europarl Corpus*. Méthodes et outils pour l'évaluation des analyseurs syntaxiques Journée ATALA, stranice 4–9, 2004.

[7] Bick, E.: *A constraint grammar parser for Spanish*. Proceedings of TIL, 2006.

[8] Bick, E.: *Degrees of Orality in Speech-like Corpora: Comparative Annotation of Chat and E-mail Corpora*. U *Proc. of the 24th Pacific Asia Conference on Language, Information and Computation. Waseda University, Sendai, Japan*, stranice 721–729, 2010.

[9] Brants, T.: *TnT: a statistical part-of-speech tagger*. U *Proceedings of the sixth conference on Applied natural language processing*, stranice 224–231. Association for Computational Linguistics, 2000.

[10] Brill, E.: *A simple rule-based part of speech tagger*. U *Proceedings of the workshop on Speech and Natural Language*, stranice 112–116. Association for Computational Linguistics, 1992.

[11] Delic, V., M. Sečujski i A. Kupusinac: *Transformation-based part-of-speech tagging for Serbian language*. U *Proceedings of the 8th WSEAS International Conference on Computational intelligence, man-machine systems and cybernetics*, stranice 98–103. World Scientific and Engineering Academy and Society (WSEAS), 2009.

[12] Dostert, L.E.: *The Georgetown-IBM experiment*. 1955). Machine translation of languages. John Wiley & Sons, New York, stranice 124–135, 1955.

[13] Erjavec, T.: *MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora*. U *Fourth Int. Conference on Language Resources and Evaluation, LREC*, svezak 4, stranice 1535–1538, 2004.

[14] fl, Automatic Language Processing Advisory Committee m.: *Language and Machines-Computers in Translation and Linguistics*. National Academy of Sciences, National, Research Council, 1966.

[15] Forcada, M.L., F.M. Tyers i G. Ramírez-Sánchez: *The Apertium machine translation platform: five years on*. U *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, stranice 3–10, 2009.

[16] Green Jr, B.F., A.K. Wolf, C. Chomsky i K. Laughery: *Baseball: an automatic question-answerer*. U *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, stranice 219–224. ACM, 1961.

[17] Johannessen, J.B., K. Hagen i A. Nøklestad: *A Constraint-Based Tagger for Norwegian*. U *17th Scandinavian Conference of Linguistics, Odense Working Papers in Language and Communication*, svezak 19, stranice 31–47, 2000.

[18] Joscelyne, A.: *John Chandioux's GramR*, 1992.

[19] Karlsson, F.: *Constraint Grammar: A language-independent system for parsing unrestricted text*, svezak 4. Walter de Gruyter, 1995.

[20] Osmann, V.: *Označavanje vrste riječi u tekstovima na hrvatskome jeziku*. 2011.

[21] Panov, D. J.: *Perevodnaja mašina P. P. Trojanskogo: sbornik materialov o perevodnoj mašine dlja perevoda s odnogo jazyka na drugie, predložennoj P. P. Trojanskim v 1933 g*. izd-vo Akademii nauk SSSR, 1959.

[22] Preradović, N.M., D. Boras i S. Kišiček: *CROVALLEX: Croatian Verb Valence Lexicon*. U *Information Technology Interfaces, 2009. ITI'09. Proceedings of the ITI 2009 31st International Conference on*, stranice 533–538, 2009.

[23] Sparck Jones, K.: *Natural language processing: a historical review*. Current Issues in Computational Linguistics: in Honour of Don Walker (Ed Zampolli, Calzolari and Palmer), Amsterdam: Kluwer, 1994.

[24] Turing, A.M.: *Computing machinery and intelligence*. Mind, 59(236):433–460, 1950.

[25] Šnajder, J., Bojana B. Dalbelo Bašić i M. Tadić: *Automatic Acquisition of Inflectional Lexica for Morphological Normalisation*. Information Processing & Management, 44(5):1720–1731, 2008.

[26] Weaver, W.: *Translation*. Machine translation of languages, 14:15–23, 1955.

[27] Weizenbaum, J.: *ELIZA—a computer program for the study of natural language communication between man and machine*. Communications of the ACM, 9(1):36–45, 1966.

[28] Woods, W.A., R.M. Kaplan i B. Nash-Webber: *The lunar sciences natural language information system*. 1972.

Sažetak

U ovome radu opisana je problematika označavanja vrste riječi, dan je pregled standardnih metoda i opisan razvoj morfosintaktičkog označivača na temelju gramatike ograničenja, a razmotren je i problem označavanja na razini sintaktičkih funkcija. Gramatika ograničenja (engl. *Constraint Grammar, CG*) jest model koji koristi kontekstno ovisna ručno pisana pravila za razrješavanje gramatičkih višeznačnosti u tekstu. Označivač na temelju gramatike ograničenja koristi morfološki leksikon dobiven poluautomatskom akvizicijom iz neoznačenog korpusa, opširan skup oznaka zasnovan na normi MULTEXT-East, valencijski leksikon hrvatskih glagola i leksikon glagola iz jezičnog para `apertium-sh-mk`, iz sustava Apertium. Gramatika sadrži 290 pravila, koja su organizirana u odjeljke za čišćenje, razrješavanje morfosintaktičkih višeznačnosti i heuristike. Gramatika je implementirana u formalizmu CG3 i prevedena *open-source* prevodiocem `vislcg3`. Preliminarni rezultati označivača iznose P: 96.1%, R: 99.8% za označavanje vrste riječi i P: 88.2%, R: 98.1% za morfosintaktičko označavanje.

Ključne riječi: CG, Gramatika ograničenja, vrste riječi, označavanje vrste riječi, hrvatski jezik, obrada prirodnog jezika, računalna lingvistika, morfosintaktičko označavanje, sintaktičko označavanje, razrješavanje morfoloških višeznačnosti

Summary

This thesis gives a description of the tasks of *part-of-speech* and morphosyntactic tagging, gives an overview of standard methods used, and describes the development of a Constraint Grammar-based morphological tagger for the Croatian language. A brief discussion on syntactic function tagging for Croatian is given as well. A Constraint Grammar (CG) uses context-dependent hand-crafted rules to disambiguate the possible grammatical readings of words in running text. The CG tagger uses a morphological analyzer based on an automatically acquired inflectional lexicon and an elaborate tagset based on MULTEXT-East, the Croatian Verb Valence Lexicon, and the verb lexicon from the Apertium language pair `apertium-sh-mk`. The grammar consists of 290 rules, organized into cleanup and mapping rules, disambiguation rules, and heuristic rules. The grammar is implemented in the CG3 formalism and compiled with the vislcg3 open-source compiler. The preliminary tagging performance is P: 96.1%, R: 99.8% for POS tagging and P: 88.2%, R: 98.1% for complete morphosyntactic tagging.

Keywords: CG, Constraint Grammar, POS, part-of-speech tagging, Croatian, natural language processing, computational linguistic, morphosyntactic tagging, syntactic tagging, morphological disambiguation

Životopis

Hrvoje Peradin, rođen u Zagrebu 07.10.1984.
Školovanje:

- 2003. - 2012. Student na PMF–MO na smjeru Računarstvo i Matematika
- 2003. - 2010. Student na PMF-–MO, završen studij prvostupnika Matematike
- 1999. - 2003. XI Gimnazija (opća), Zagreb;

Radna iskustva vezana uz studij:

- Google Summer of Code 2012. rad na jezičnom paru
  Serbo-Croatian – Slovene (`apertium-sh-sl`)
- Mentor na Google-Code-In 2011.-12.
- Google Summer of Code 2011. rad na jezičnom paru
  Serbo-Croatian – Macedonian (`apertium-sh-mk`)

Publikacije:

- "A rule-based machine translation system from Serbo-Croatian to Macedonian", H. Peradin, F. Tyers, FreeRBMT12, Third International Workshop on Free/Open-source Rule-based Machine Translation
- "Towards a Constraint Grammar Based Morphological Tagger for Croatian", H. Peradin, J. Šnajder, TSD 2012, 15th International Conference on Text, Speech and Dialogue