Bojana P. Đorđević[*]
Centar za srpski jezik i kulturu Azbukum

УДК
811.163.41'36:004
ДОИ
https://doi.org/10.18485/analiff.2017.29.1.4

SRPTAG: LEKSIKALIZOVANA GRAMATIKA ADJUNGOVANIH STABALA ZA SRPSKI JEZIK

U ovom radu sažeto predstavljamo trenutno stanje prve formalne gramatike srpskog jezika – SrpTAG, napravljene sa ciljem da se opišu osnovne rečenične strukture srpskog jezika, a za potrebe automatske sintaksičke analize. Kroz uvod u temu automatske obrade teksta, čitaoca u drugom poglavlju uvodimo u formalizam FBLTAG, formalizam leksikalizovane gramatike stabala na kome je gramatika bazirana. U trećem poglavlju detaljno predstavljamo osnovne izbore napravljene pri formiranju SrpTAG – počevši od tradicionalne gramatike koja čini polazište u našem radu, izbora kategorija, obeležja i funkcija, pa do samog definisanja FBLTAG stabala za srpski jezik. Ukratko predstavljamo i leksikon, koji čini nerazdvojni deo ove gramatike. Broj i sastav jedinica koje je trenutno moguće automatski prepoznati dajemo u delu o evaluaciji gramatike, dok na samom kraju rada postavljamo perspektive za njen dalji razvoj.

Ključne reči: SrpTAG, FBLTAG, formalna gramatika, parsiranje, automatska sintaksička analiza, obrada prirodnih jezika

1. Uvod

Potreba da se jezička obrada prepusti računaru prvi put se javila paralelno sa prvim računarima – sredinom prošlog veka. Ti napori su u početku bili usmereni na automatsko tj. mašinsko prevođenje i bili su direktno vezani za tadašnje političke prilike u svetu – potrebu da se u vreme hladnog rata na brz način i automatski prevodi između ruskog i engleskog. Neuspeh koji je ovaj desetogodišnji projekat doživeo preusmerio je pažnju naučnika sa mašinskog prevođenja na druge sfere u kojima su zajednički napori programera i lingvista mogli da daju značajnije rezultate. Oblast automatske obrade jezika je danas veoma razgranata i sastoji se od velikog broja postupaka za obradu jezika i različitih disciplina, među kojima su

[*] bojana@lingvistika.org

pronalaženje informacija, ekstrakcija informacija, odgovaranje na pitanja, sumarizacija teksta, ekstrakcija termina, automatsko indeksiranje teksta, učenje jezika uz pomoć računara, prepoznavanje i generisanje govora, mašinsko učenje **i već pomenuto mašinsko prevođenje**.

Svaki od pomenutih pristupa zavisi, u manjoj ili većoj meri, od resursa koji su razvijeni za dati jezik, a od kojih su centralni korpus – kolekcija tekstova u elektronskom obliku nad kojom se mogu vršiti analize, elektronski rečnik – rečnik u mašinski čitljivom obliku koji se koristi za automatsku morfološku analizu teksta, i gramatika – formalna gramatika koja se koristi za automatsko prepoznavanje strukture rečenica.

Za srpski jezik prva dva resursa postoje i nadograđuju se već neko vreme. Trenutno najveći korpus savremenog srpskog jezika, SrpKor (Krstev & Vitas, 2005; Utvić, 2013), razvijen je na Matematičkom fakulteta u Beogradu i njegova veličina je 122 miliona reči. Rad na morfološkim elektronskim rečnicima takođe je započet na Matematičkom fakultetu u Beogradu (Krstev, 2008), a njihovo dopunjavanje neprekidno traje. Rečnik se trenutno sastoji od 140.000 prostih i 18.000 višečlanih reči. Tema ovog rada je prva formalna gramatika srpskog jezika SrpTAG, razvijena u okviru doktorske teze autorke (Đorđević, 2017).

Formalna gramatika srpskog jezika SrpTAG napravljena je sa ciljem automatskog prepoznavanja osnovnih i određenog broja izvedenih rečeničnih struktura. Potrebno je napomenuti da iako ova gramatika jeste čitljiva, ona je pre svega namenjena računarskoj obradi teksta, odnosno namenjena je računaru. Ovako formirana SrpTAG se, kroz nekoliko faza transformacije, koristi unutar parsera – programa za automatsku sintaksičku analizu – i omogućava automatsku analizu pojedinačnih rečenica.

2 O formalizmu

Formalizam na kome smo zasnovali SrpTAG jeste leksikalizovana gramatika adjungovanih stabala zasnovana na obeležjima – FBLTAG (Vijay-Shanker & Joshi 1988; Vijay-Shanker & Joshi 1991). U osnovi ovog formalizma nalazi se leksikalizovana gramatika adjungovanih stabala – LTAG (Schabes, Abeillé, & Joshi, 1988; Schabes, 1990) koja je dodatno obogaćena obeležjima. Ovde ćemo najpre ukratko predstaviti osnovne postavke gramatike LTAG (poglavlje 2.1), a segment obeležja ćemo dodati u nastavku rada (poglavlje 2.2).

2.1 LTAG

LTAG je leksikalizovana generativna gramatika koja za svoju osnovnu jedinicu ima stablo. Činjenica da je leksikalizovana znači da je svaka struktura koju ova gramatika opisuje direktno vezana za leksemu koja je nosilac date strukture. To da je generativna znači da opisuje pravila na osnovu kojih je moguće automatski generisati strukture jezika, a te strukture su u ovom slučaju opisane isključivo kao stabla.

2.1.1 Tipovi stabala

LTAG pravi razliku između dva tipa stabala: inicijalnih i pomoćnih stabala (Slika 1). Inicijalna stabla[1] odgovaraju minimalnim nerekurzivnim jezičkim strukturama i predstavljaju argumentsku strukturu jedinice koja je njihovo sidro. Sidro je čvor preko koga je stablo povezano sa leksikonom. Za stablo čije je sidro glagol u funkciji predikata, inicijalno stablo se sastoji od svih obaveznih argumenata datog glagola. U primeru na slici 1 dato je stablo glagola *spavati* (stablo α_1), koji kao neprelazan glagol u svojoj argumentskoj strukturi ima samo subjekat, te je za njega rezervisana pozicija u stablu ovog glagola.

Pomoćna stabla[2] odgovaraju minimalnim rekurzivnim jezičkim strukturama i najčešće se koriste kao modifikatori. Na slici 1 dato je stablo prideva *lep* (stablo β_1). Pomoćna i inicijalna stabla se zajednički nazivaju elementarnim stablima.

2.1.2 Operacije

LTAG propisuje dve operacije: zamenu i pripajanje. Pripajanje je ključna operacija LTAG, kako ovoj gramatici obezbeđuje rekurzivnost. Pripajanje podrazumeva spajanje pomoćnog stabla i bilo kog drugog stabla, uključujući i pomoćno. Slika 2 daje prikaz operacije pripajanja na srpskom jeziku.

Zamena predstavlja proizvod leksikalizacije i kao takva, zadužena je za unošenje leksičkih elemenata u stablo. Može se odvijati samo na neterminalnim čvorovima na granici stabala, koji po konvenciji nose oznaku ↓. Prikaz operacije zamene na primerima iz srpskog jezika dat je niže (Slika 3).

Stabla nastala primenom bilo koje od ovih operacija nazivaju se izvedenim stablima[3].

---

1 Po konvenciji se obeležavaju simbolom α.

2 Po konvenciji se obeležavaju simbolom β.

3 Stabla nastala primenom zamene i pripajanja nazivaju se izvedena stabla i po konvenciji obeležavaju simbolom γ.

2.2 FBLTAG

FBLTAG, formalizam koji ćemo nadalje koristiti u radu, na svaki čvor stabala koje smo ovde opisali dodaje obeležja. Ta obeležja su karakteristična po tome što su dupla, te svaki čvor ima gornja (t) i donja obeležja (b) (Slika 4). Gornja obeležja pokazuju odnos datog čvora prema nadstablu, i predstavljaju pogled na čvor odozgo. Donja obeležja pokazuju odnos prema podređenim čvorovima, i predstavljaju pogled na čvor odozdo.

Ipak, osnovni razlog za primenu duple strukture obeležja leži u prirodi operacije pripajanja. Ono što se prilikom pripajanja dešava jeste da se čvor na kome se vrši pripajanje (Slika 5, čvor $X^{t:f}_{b:g}$) deli na dva dela – jedan koji se spaja sa korenim čvorom pomoćnog stabla ($X^{t:f1}_{b:g1}$) i drugi koji se spaja sa priključnim čvorom pomoćnog stabla ($X^{*\,t:f2}_{b:g2}$). Dvostruka struktura obeležja na tom čvoru omogućava mu ovakvo deljenje. Tom prilikom se gornje obeležje čvora na kome se vrši priključivanje (t:f) spaja sa gornjim obeležjem korenog čvora pomoćnog stabla ($t:f_1$), a donje obeležje čvora na kome se vrši priključivanje (b:g) spaja sa donjim obeležjem priključnog čvora ($b:g_2$). Da bi do spajanja moglo da dođe, potrebno je proveriti da li su vrednosti navedenih obeležja međusobno kompatibilne, odnosno iste. Ukoliko jesu, dolazi do njihovog stapanja u jedinstveno obeležje, što je proces koji se naziva unifikacija.

Slika 6 daje primer koji smo videli ranije (Slika 2), ali ovoga puta s pridodatim obeležjima na relevantnim čvorovima, kako bi se mogao videti proces unifikacije.

2.3 Prikladnost (FB)LTAG za primenu na srpski jezik

LTAG, kao i njegova verzija FBLTAG, ima tri obeležja koja je čine prikladnom za automatsku obradu jezika, posebno za jezike sa slobodnim redom reči kakav je srpski:

- Ovaj formalizam karakteriše **proširen domen lokalnosti**. Ovo svojstvo proizlazi iz same prirode LTAG kao formalizma koji za osnovnu jedinicu ima elementarno stablo. Dobro formirano predikatsko elementarno stablo zahteva da se unutar njega nalaze svi argumenti predikata. Prateći ovaj princip, u (FB)LTAG se rečenice sa npr. izmeštenim objektom *na koga* u rečenici *Na koga Marko misli.*

definišu kao jedno od primarnih elementarnih stabala (a ne izvedenih), čime se postiže da odnosi među svim argumentima i dalje budu definisani lokalno[4]. Iz ovog razloga u TAG ne postoji potreba za operacijom transformacije koja bi generisala izmenjene niske i tražila način da se izmeštenoj niski dodele odgovarajuća obeležja.

- **Rekurzija** u LTAG gramatikama čuva domen zavisnosti. Rekurzija u TAG obezbeđuje se primenom operacije pripajanja. Pripajanje se vrši na elementarnom stablu (lokalnom domenu zavisnosti) u koje se na taj način unosi nov element. Budući da je zavisnost već jasno definisana među čvorovima elementarnog stabla, unošenje novog stabla ni na koji način ne menja njihov odnos – skup njihovih deljenih obeležja i dalje je onaj iz početne elementarne strukture. Na primer, rečenica Šta Marko misli da Ivan voli, u kojoj se argument šta tipično smatra izmeštenim iz zavisne rečenice *Ivan voli*, u (FB)LTAG se rekonstruiše kao elementarna rečenica Šta Ivan voli – budući da svi argumenti glagola uvek moraju biti zajedno unutar elementarne rečenice – u koju je zatim pripajanjem uneta „glavna" rečenica *Marko misli da*. U tom smislu se zapravo i ne može govoriti o udaljenosti argumenata, koji su i pored linearne udaljenosti uvek lokalno vezani.
- Zahtev LTAG gramatika da svako stablo bude povezano sa leksikonom čini ovaj formalizam u potpunosti **leksikalizovanim**. Imajući u vidu da je u ovim gramatikama leksema dominantni nosilac sintaksičke informacije, osnovna struktura u koju može ući leksema se već unutar leksikona povezuje sa svim drugim strukturama u kojima se ona može naći. Ovo povezivanje se vrši pomoću leksičkih pravila. Uzmimo za primer glagol *kupovati*. Osnovna struktura u koju ulazi ovaj glagol je subjekat-predikat-pravi objekat (Žena kupuje kafu.). U upotrebi, ovaj glagol se može javiti u različitom tipu alternativnih struktura, kao što je recimo upitna (Šta žena *kupuje?*), pasivna, u obliku participskog pasiva (*Kafa je kupovana.*), ili pasivna, u obliku refleksivnog pasiva (*Kafa se kupuje.*). Ovom prilikom može biti promenjen oblik glagola (participski pasiv), neki od argumenata mogu biti izostavljeni (participski pasiv i pitanje), može im biti promenje-

4 Ovo stablo je sa osnovnom verzijom stabla *Marko misli na Milenu* povezano leksičkim pravilom. Za opis leksičkih pravila, videti stavku o leksikalizovanosti.

na funkcija (participski pasiv) ili pak mogu biti potpuno uklonjeni iz konstrukcije (refleksivni pasiv). U ovom slučaju definišu se tri leksička pravila: za participski pasiv, za refleksivni pasiv i za pitanje, koja imaju dvojaku ulogu. S jedne strane, ona povezuju sva stabla u koja ulazi glagol *kupovati* sa osnovnim stablom ovog glagola, a s druge strane, na osnovu njih se proverava da li određena struktura zadovoljava kriterijum da bude povezana u skup struktura jedne lekseme. Sva stabla koja su međusobno povezana leksičkim pravilima organizovana su u porodicu stabala.

Ovakva organizacija leksikona omogućava da se napravi razlika između leksema koje pripadaju istoj osnovnoj rečeničnoj porodici, ali nemaju iste alternativne strukture, kao u slučaju glagola *voleti* i *znati*, od kojih drugi nema odlik participskog pasiva:

$$\text{voleti}\begin{bmatrix} \text{n0Vn1dir} \\ \text{ppasiv} \quad + \\ \text{refpasiv} \quad + \end{bmatrix}\begin{bmatrix} \text{n0Vn1dir} \\ \text{ppasiv} \quad + \\ \text{refpasiv} \quad + \end{bmatrix}\text{znati}\begin{bmatrix} \text{n0Vn1dir} \\ \text{ppasiv} \quad - \\ \text{refpasiv} \quad + \end{bmatrix}\begin{bmatrix} \text{n0Vn1dir} \\ \text{ppasiv} \quad - \\ \text{refpasiv} \quad + \end{bmatrix}$$

3 SrpTAG

3.1 Izbor tradicionalne gramatike

Budući da je SrpTAG prva formalna gramatika srpskog jezika, odlučili smo da se u poslu njene izrade oslonimo na neku od gramatika namenjenih ljudskoj upotrebi, i pored toga što se one za ove potrebe generalno smatraju nezadovoljavajućim (Erbach & Uszkoreit, 1990). Naš izbor je u ovom slučaju bila *Gramatka srpskog jezika – udžbenik za I, II, III i IV razred srednje škole* (Stanojčić & Popović, 1997). I pored toga što pristup LTAG nije uvek kompatibilan sa pristupom unutar ove gramatike, njena klasifikacija osnovnih rečeničnih konstituenata i rečeničnih modela čini našu polazišnu tačku pri definisanju osnovnih stabala gramatike, leksičkih pravila i porodica stabala.

Kao uzor pri izradi SrpTAG koristili smo opsežnu FBLTAG napravljenu za francuski jezik, predstavljenu u celosti u (Abeillé, 2002).

3.2 Izbor morfosintaksičkog opisa

Kao polazište za morfosintaksički opis (MSO) gramatike SrpTAG, uzeli smo MSO koji se koristi u elektronskim rečnicima srpskog jezika, opisanim u (Krstev, 2008). Kako je ovaj rečnik već prilagođen za potrebe automatske obrade teksta, format njegovih obeležja bio je sasvim prikladan za MSO naše gramatike. Ovoj grupi obeležja dodali smo obeležja koja su proizašla iz potreba same formalne gramatike. U SrpTAG tako postoji 13 terminalnih kategorija i 7 frazalnih kategorija.

Kategorije **N** – imenice, **V** – glagoli, **A** – pridevi, **ADV** – prilozi, **PREP** – predlozi, **NUM** – brojne reči, **PAR** – rečce, **PRO** – zamenice i **INT** – uzvici preuzete su direktno i u neizmenjenom obliku iz elektronskih rečnika. Kategorija PRO, koja se u elektronskom rečniku koristi za sve tipove zamenica, u SrpTAG je podeljena na PRO, što je oznaka koju koristimo za sve pune oblike imenica, i **Cl**, što je oznaka koju koristimo za enklitičke oblike. Kategorija CONJ, koja postoji u elektronskom rečniku i koristi se za sve tipove veznika, u SrpTAG je razdvojena na dve grupe – **CONJ** za naporedne veznike, i **C** za subordinativne veznike. Pored ove dve dodatne kategorije, sasvim nova terminalna kategorija je **NEG** koju koristimo za označavanje pozitivne ili negativne vrednosti strukture.

Za gramatiku koristimo i sedam frazalnih kategorija: **NP** – imeničke fraze, **VP** – glagolske fraze tj. predikati, **AP** – pridevske fraze, **ADVP** – priloške fraze, **PP** – predloške fraze, **NUMP** – brojne fraze i **S** – zavisne i nezavisne rečenice. Minimalna glagolska fraza u SrpTAG sastoji se od glagola, a proširena od refleksivne rečce *se*, negacije u obliku rečce *ne*, pomoćnog glagola i modalnog glagola. Predikativ ne ulazi u sastav glagolske fraze.

3.3 Izbor obeležja

Veliki broj obeležja koje koristimo u radu poklapa se sa obeležjima koja se javljaju u elektronskim rečnicima srpskog jezika. Kao i kod kategorija, i ovde je jedan broj obeležja dodat za konkretne potrebe SrpTAG. Sva obeležja možemo podeliti na morfosintaksička, sintaksička i obeležja koja ograničavaju primenu leksičkih pravila.

- U morfosintaksička obeležja ulaze: **gen** – rod, **num** – broj, **case** – padež, **def** – određenost, **pers** – lice, **ord** – redni broj, **asp** – vid glagola, **mod** – modus glagola, **form** – glagolska vremena i načini.

- Od sintaksičkih obeležja razlikujemo: **neg** – negacija, **aux** – pomoćni glagol, **wh** – upitni oblik, **ref** – refleksivni, **func** – funkcija, **subg** – slaganje sa subjektom u rodu, **subp** – slaganje sa subjektom u licu, **subn** – slaganje sa subjektom u broju, **modal** – modalni glagol, **fazni** – fazni glagol, **cop** – kopulativni glagol.

Oznake subg, subn i subp koristimo da naznačimo da izostavljeni subjekat zavisne rečenice deli kategorije roda, broja i lica sa subjektom glavne rečenice. Čvorovi stabala u SrpTAG su obeleženi kategorijama, te se funkcija date jedinice uvek navodi u okviru obeležja func.

- U treću grupu obeležja spadaju obeležja **pasiv**, kojim se naznačava da li glagol može ući u konstrukciju s participskim pasivom, i oznake **bez1** i **bez2**, kojima se označava da prvi, odnosno drugi argument glagola može biti izostavljen iz rečenične strukture (stabla).

Dok se prve dve grupe obeležja javljaju paralelno uz stabla i u leksikonu, treća grupa obeležja se javlja samo u leksikonu, gde ograničava broj i tip stabala u kojima se određeni glagol može naći.

U ovoj fazi SrpTAG ne postoji nijedno semantičko obeležje, iako ona postoje u elektronskim rečnicima.

3.4 Izbor i realizacija gramatičkih funkcija

Funkcije od kojih smo krenuli u radu jesu mahom one koje su navedene kao funkcije koje ulaze u rečenične modele u gramatici (Stanojčić & Popović, 1997). U pitanju su subjekat i logički subjekat, pravi objekat, nepravi objekat i priloška dopuna. Predikativ se u okviru LTAG ne smatra argumentom, već nosiocem stabla (sidrom), svojevrsnim predikatom u neglagolskom obliku, te iz tog razloga nije uključen u ovaj popis funkcija. U analizi ne uzimamo u obzir dopunski predikativ kao jedinicu, kako nije bilo sasvim jasno kako je obraditi unutar LTAG sistema. Odredbe trenutno nisu obrađene unutar SrpTAG.

Pomenute gramatičke funkcije mogu biti realizovane kao pojedinačne kategorije (imenice, zamenice, prilozi), sintagme ili fraze (imenička, priloška, predloška ili brojna) i kao zavisne rečenice ili rečenice s glagolom u infinitivu.

3.5 Leksikon

Leksikon SrpTAG se sastoji od morfološkog i sintaksičkog leksikona. Morfološki leksikon sadrži lemu[5], njenu konkretnu formu i njena morfosintaksička i sintaksička obeležja. Sintaksički leksikon sadrži lemu i naziv porodice stabala kojoj data lema pripada. U sintaksičkom leksikonu se kao obeležja mogu javiti obeležja koja ograničavaju primenu leksičkih pravila, odnosno broj stabala unutar navedene porodice u koju data lema može ući.

I sintaksički i morfološki leksikon su trenutno ručno pravljeni i sastoje se od svega 57 lema i oko stotinu njihovih oblika. Ova veličina leksikona se koristi samo u svrhe bazičnih testiranja gramatike. U planu je rad na konverziji elektronskih rečnika srpskog jezika u ovaj format rečnika, čime će se u vrlo kratkom roku veličina oba rečnika znatno uvećati.

3.6 Elementarna stabla

SrpTAG se trenutno sastoji od 203 elementarna stabla. Ovaj broj ne predstavlja pun opseg gramatike koja je trenutno teorijski opisana i proširivaće se u budućnosti, kako se i leksikon i broj formalnih opisa stabala bude uvećavao. Takođe, u ovaj broj nisu uključene porodice stabala s rečeničnim argumentima. Među ovim stablima se nalaze ona čija su sidra glagoli, ali i ona čija su sidra imenice, zamenice, pridevi i prilozi.

3.6.1 Elementarna stabla glagola

Polazeći od rečeničnih modela subjekatsko-predikatskih i bezličnih rečenica definisanih u (Stanojčić & Popović, 1997) definisali smo ukupno 51 porodicu stabala sa glagolom kao svojim centralnim elementom, što je istovremeno i 51 supkategorizacioni okvir glagola. Pored toga, za sve porodice na koje su primenjiva, definisana su i leksička pravila za participski pasiv, refleksivni pasiv, obezličavanje i refleksivizaciju[6]. Red reči nije posebno obrađivan, sem kod refleksivnih glagola, gde su definisana dva različita reda reči – jedan kada postoji subjekat, a drugi kada je subjekat izostavljen.

5 Osnovni, rečnički oblik reči.

6 Pod refleksivizacijom podrazumevamo odnos između rečenice tipa *Marija češlja sebe.* i *Marija se češlja*., gde refleksivna rečca *se* predstavlja enklitički oblik zamenice *sebe* u funkciji pravog objekta.

3.6.2 Elementarna stabla ostalih kategorija

Trenutno među elementarnim stablima koja nisu glagolska imamo stabla s kopulativnim glagolom. Naime, konstrukcije s kopulativnim glagolom se u TAG smatraju konstrukcijama u kojima je centralni čvor (sidro) predikativ, a ne kopula ili kopula i predikativ zajedno. Kopula se u ovakve konstrukcije unosi zamenom, kao i jedinica koja npr. vrši funkciju subjekta. Razlikujemo tako pet porodica stabala koje smo definisali na osnovu subjekatsko-predikatskih rečeničnih modela iz (Stanojčić & Popović, 1997) i dve porodice stabala definisane na osnovu bezličnih modela iz iste gramatike. Za stabla sa kopulativnim glagolom definisane su posebne porodice stabala koje su orijentisane oko kategorije koja je nosilac funkcije predikativa – imenice, zamenice, prideva i neke od jedinica s priloškom funkcijom (prilog, predloško-padežna konstrukcija, brojna konstrukcija).

3.7 Primena i evaluacija

Gramatika kakvu smo definisali u prethodnim poglavljima predstavlja ulaz za parser. Parser koji koristimo u ovom procesu je parser TuLiPA (Kallmeyer et al., 2008). Pored gramatike, ulaz parsera predstavljaju i leksikona i gramatika, kao i rečenica koju treba analizirati, pri čemu je izlaz struktura date rečenice predstavljena kao rečenično stablo. Slika 7 daje izgled parsera i primer analize koja se dobija pomoću njega.

Pogledajmo kakav je trenutni stepen prepoznavanja svakog od glagolskih argumenata i samih glagola.

- **Subjekat:** Sistem u ovom trenutku prepoznaje sve imeničke i zameničke vrste subjekta – kao lične zamenice i odgovarajuće upitne zamenice za formiranje subjekatskog pitanja (*ko, šta*, kao u *Ko spava?*). Kongruencija subjekta sa pomoćnim glagolom, kopulom i participom funkcioniše ispravno kroz sve klase. Rečenični subjekat ne prepoznajemo u ovom trenutku.

Jedina pozicija subjekta koju trenutno prepoznajemo jeste ona na prvom mestu u rečenici. Izuzetak su rečenice sa logičkim subjektom, za koje smo formulisali eksperimentalna stabla koja imaju slobodan red reči, te subjekat u ovom slučaju prepoznajemo iza glagola. Nerealizovan subjekat se ispravno prepoznaje kroz modele s glagolskim predikatom, uključujući i refleksivne glagole, međutim u modelima s kopulativnim glagolima ne uzimamo u obzir to da subjekat može biti izostavljen.

- **Predikat:** Sistem ispravno prepoznaje glagolske predikate u pozitivnom i negativnom obliku u prezentu, uključujući i slučajeve sa refleksivnim glagolima (npr. *Marko se ne šali.*). Oblici refleksivnog pasiva i obezličene varijante glagola se uspešno prepoznaju u oba slučaja. Trenutno ne prepoznajemo negaciju na pomoćnim glagolima, kopuli, i u participskom pasivu. Prošlo vreme i buduće vreme prepoznajemo na glagolima koji nisu refleksivni i nisu kopule. Pomoćne glagole trenutno prepoznajemo samo kada se nalaze neposredno ispred glagola.

Ne prepoznajemo futur u obliku da+prezent i ne prepoznajemo modalne glagole – ni jedna ni druga konstrukcija trenutno nisu unete u sistem.

- **Pravi objekat:** Trenutno prepoznajemo imenički i zamenički pravi objekat, kao i varijantu sa upitnim pravim objektom (*Koga Sonja voli?*). Klasa za rečenični pravi objekat još uvek ne funkcioniše, dok pravi objekat u obliku enklitike dosad nismo obrađivali.

Pravi objekat trenutno prepoznajemo samo na poziciji iza glagola.

- **Nepravi objekat:** Uspešno prepoznajemo imeničke i zameničke objekte, uključujući i upitnu varijantu nepravog objekta (*Kome daje lutku?*). Slika 8 daje primer analize s ovakvim upitnim nepravim objektom. Prepoznajemo neprave objekte u obliku predloško-padežne konstrukcije, iako trenutno sve predloge tretiramo kao jedinice koje se unose zamenom, a ne kao dodatna sidra uz glagol. Trenutno ne prepoznajemo predloško-padežne konstrukcije u upitnom obliku (*U čemu spava Marko?*), enklitičke oblike nepravog objekta, kao ni rečenični nepravi objekat.

Nepravi objekat prepoznajemo samo na poziciji iza glagola. U rečenicama u kojima postoje i pravi i nepravi objekat, prepoznajemo oba njihova međusobna rasporeda.

- **Priloška dopuna:** Od priloških dopuna trenutno prepoznajemo priloge i predloško-padežne konstrukcije. Iako imamo razvijenu klasu za brojne dopune, trenutno nije implementirana. Prepoznajemo pitanja u kojima su prilog i predloško-padežna konstrukcija zamenjeni upitnom priloškom rečju (npr. *Gde Marko živi?*)

Priloške dopune trenutno prepoznajemo samo na poziciji iza glagola.

- **Logički subjekat:** Logički subjekat prepoznajemo u imeničkom i zameničkom obliku, i kao upitnu zamenicu (*Koga boli glava?*).

Pozicija na kojoj trenutno prepoznajemo logički subjekat jeste ispred glagola.

- **Predikativi:** Imenski predikativ prepoznajemo kao imenicu ili zamenicu. Kod pridevskih predikativa ispravno funkcioniše kongruencija sa subjektom. Od priloških predikativa trenutno prepoznajemo samo priloge. Ni za jedan od predikativa trenutno nemamo definisan upitni oblik. Za predikative trenutno nemamo definisanu klasu koja omogućava prepoznavanje rečenica kao što je *Pametan je*, ali možemo da prepoznajemo bezlične modele ovakve strukture, sa prilogom kao glagolskim delom predikata i subjektom koji nije prisutan u strukturi – *Dobro je*.

Kompiliranje gramatike ove veličine i sastava traje par sekundi preko jednog minuta (minut i tri sekunde). Dvosmislenost prilikom parsiranja je trenutno minimalna, i slobodno se može reći nepostojeća. Računamo na to da se dvosmilenost u određenoj meri povećati kada se u leksikone unese veći broj jedinica i implementira veći broj stabala.

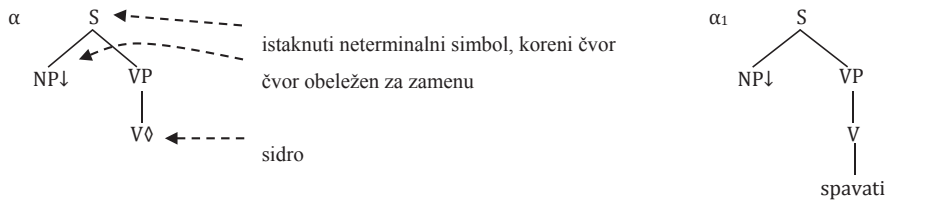
4 Zaključak i dalji rad

U ovom radu smo dali sažet prikaz prve formalne gramatike srpskog jezika SrpTAG. Dali smo prikaz formalizma na kome se ova gramatika bazira i opis trenutnog stanja gramatike. Naš cilj prilikom izrade ove gramatike bio je da obradimo osnovne strukture srpskog jezika definisane kroz rečenične modele u gramatici (Stanojčić & Popović, 1997) i te osnove su postavljene.

Primeri koje trenutno nije moguće prepoznati mogu se podeliti na dve grupe. U jednu grupu spadaju oni za koje ne postoji tehnička ni jezička poteškoća i koji samo još uvek nisu obrađeni u sistemu. U ovu grupu spadaju negacija i prošlo i buduće vreme za kopulativne glagole, prepoznavanje primera bez subjekta za kopulativne glagole, prepoznavanje brojnih fraza, modalnih glagola i sl.. Ovo su prvi zadaci koji predstoje u našem daljem bavljenju ovom temom. U drugu grupu spadaju pojave koje izazivaju neku vrstu tehničke ili jezičke poteškoće, a za koje će biti potrebno značajnije modifikovati sistem. U ovu grupu spada obrada predloško-padežnih konstrukcija kao upitnih fraza i kao predikativa, tretman rečeničnih argumenata, tretman enklitika i tretman reda reči. Ovi zadaci spadaju u dugoročnije planove na unapređivanju SrpTAG.

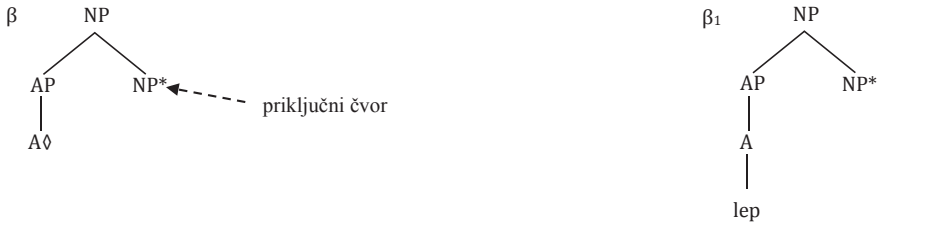
INICIJALNO STABLO

Shema inicijalnog stabla ***Primer inicijalnog stabla***

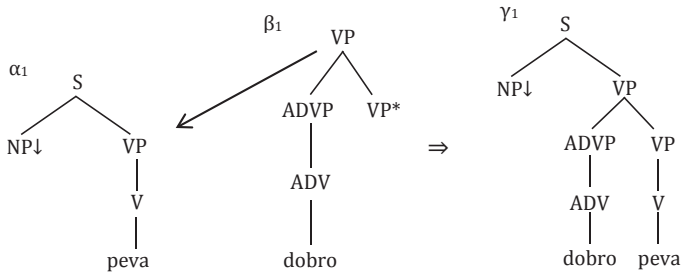


POMOĆNO STABLO

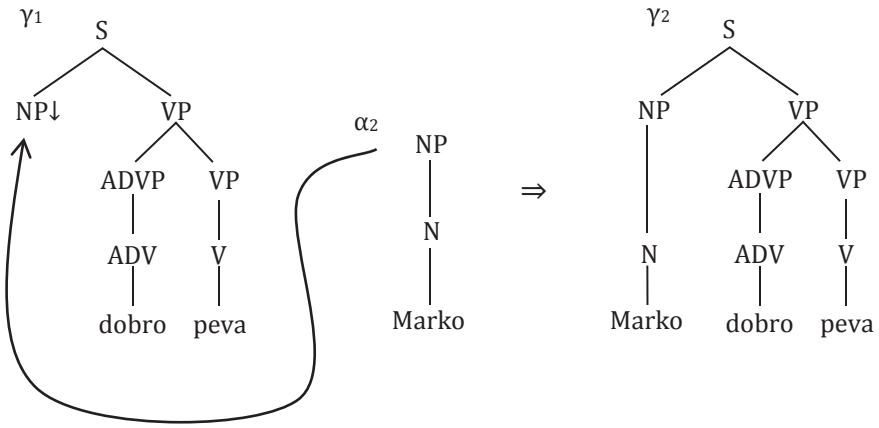
Shema pomoćnog stabla ***Primer pomoćnog stabla***



Slika 1 Sheme inicijalnih i pomoćnih stabala i njihovi primeri za srpski jezik



Slika 2 Primer operacije pripajanja na srpskom jeziku



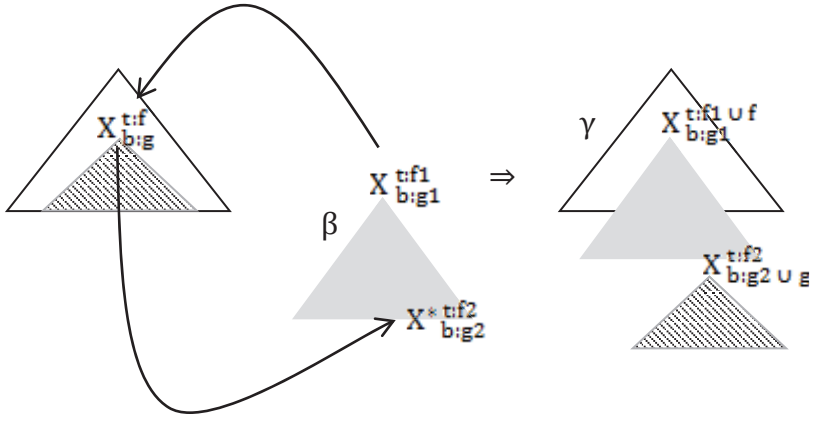
Slika 3 Primer operacije zamene na srpskom jeziku

NP

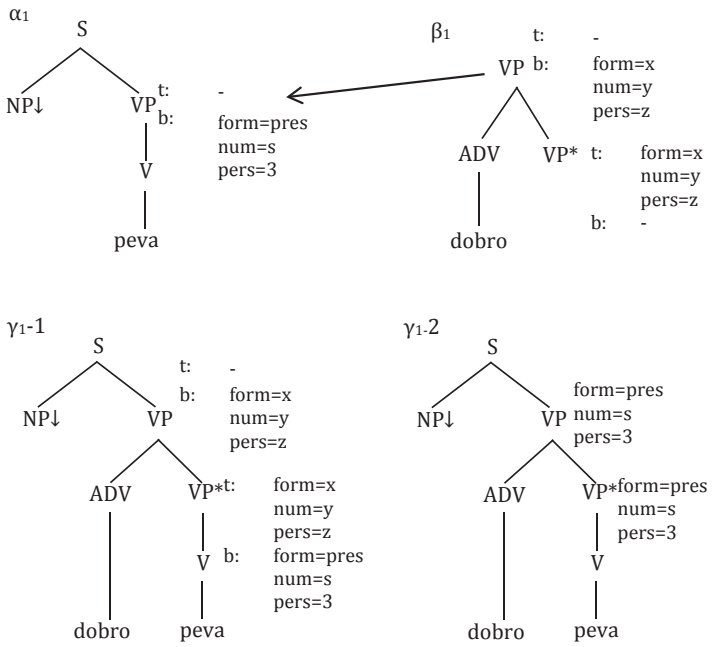
t: num=x
gen=y
case=z

b: num=x
gen=y
case=z

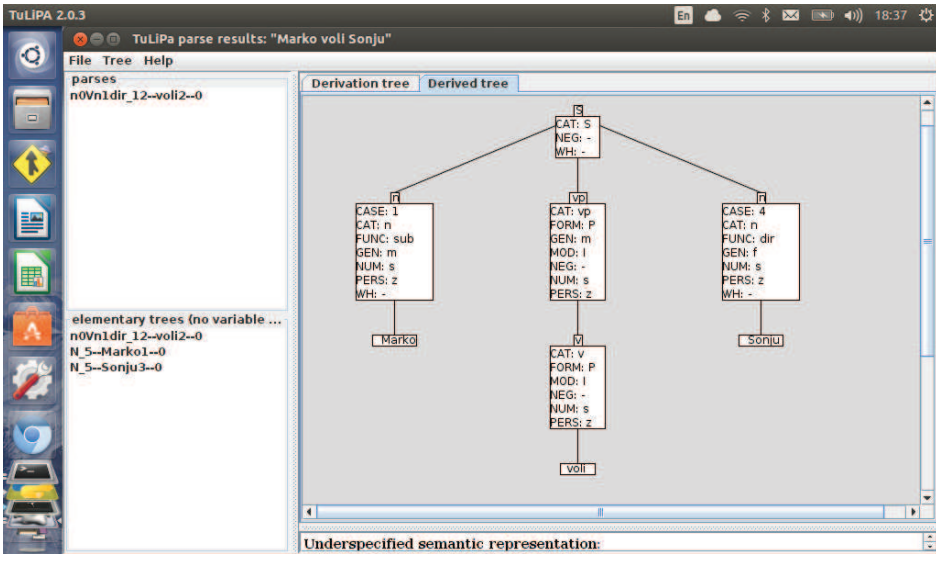
Slika 4 Shematski prikaz obeležja na čvoru imeničke fraze



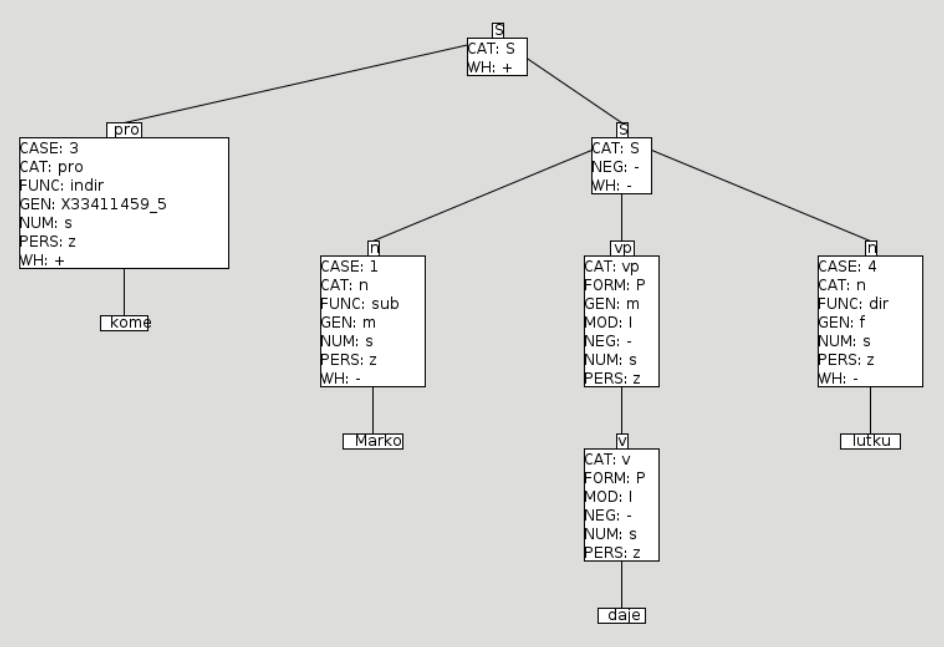
Slika 5 Shematski prikaz unifikacije obeležja pri pripajanju



Slika 6 Primer unifikacije pri pripajanju na primeru srpskog jezika



Slika 7 Rezultat parsiranja unutar parsera TuLiPA za rečenicu *Marko voli Sonju*



Slika 8 Izlaz parsera TuLiPA za rečenicu *Kome Marko daje lutku?*

Literatura

Abeillé, A. (2002). *Une grammaire électronique du français*. Paris: CNRS.

Đorđević, B. (2017). *Izrada osnova formalne gramatike srpskog jezika upotrebnom metagramatike*. (neobjavljena doktorska disertacija). Filološki fakultet, Beograd.

Erbach, G., & Uszkoreit, H. (1990). *Grammar Engineering: Problems and Prospects*.

Kallmeyer, L., Lichte, T., Maier, W., Parmentier, Y., Dellert, J., & Evang, K. (2008). TuLiPA : Towards a Multi-Formalism Parsing Environment for Grammar Engineering. *Proceedings of the Workshop on Grammar Engineering Across Frameworks (Coling 2008)*, (August), 1–8.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Belgrade: Faculty of Philology, University of Belgrade.

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon - Mutual Incompletness. In P. Danielsson & M. Wagenmakers (Eds.), *Proceedings of the Corpus Linguistics Conference*. Birmingham.

Schabes, Y. (1990). *Mathematical and computational aspects of lexicalized grammars*. University of Pennsylvania.

Schabes, Y., Abeillé, A., & Joshi, A. K. (1988). Parsing Strategies with "Lexicalized" Grammars: Application to Tree Adjoining Grammars. *Proceedings of the 12th Conference on Computational Linguistics*, 578–583.

Stanojčić, Ž., & Popović, L. (1997). *Gramatika srpskoga jezika: udžbenik za I, II, III i IV razred srednje škole*. Beograd: Zavod za udžbenike i nastavna sredstva.

Utvić, M. (2013). *Izgradnja referentnog korpusa savremenog srpskog jezika*. (neobjavljena doktorska disertacija). Filološki fakultet, Beograd.

Vijay-Shanker, K., & Joshi, A. K. (1988). Feature Structures Based Tree Adjoining Grammar. *Proceedings of COLING*, (October), 714–719.

Vijay-Shanker, K., & Joshi, A. K. (1991). *Unification-Based Tree Adjoining Grammars*. *Technical Reports (CIS)*.

Bojana Đorđević

Summary

SRPTAG: A LEXICALIZED TREE-ADJOINING GRAMMAR FOR SERBIAN

This paper aims to present the first formal grammar for parsing Serbian named SrpTAG. Through the introduction into the world of automatic processing of text, it presents the basics of the underlying formalism used to build SrpTAG – FBLTAG. This tree-based grammar was used to mold the core sentence models presented in a widely-used traditional grammar of Serbian language and together with other choices presented in the paper served as the basis for the formal grammar of Serbian. The evaluation part brings a detailed presentation of the tasks performed and the ones that are yet to be completed, which are further divided into the short-term and long-term ones in the conclusion.

Key words: SrpTAG, FBLTAG, formal grammar, parsing, automated syntactic analysis, natural language processing