Označevanje udeleženskih vlog v učnem korpusu za slovenščino

Simon Krek,*+ Polona Gantar,♦ Kaja Dobrovoljc,† Iza Škrjanec‡

* Laboratorij za umetno inteligenco, Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana
+ Center za jezikovne vire in tehnologije Univerze v Ljubljani, Večna pot 113, 1000 Ljubljana
simon.krek@ijs.si
♦ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
apolonija.gantar@ff-uni.lj.si
† Zavod za uporabno slovenistiko, Trojina, Trg republike 3, 1000 Ljubljana
kaja.dobrovoljc@trojina.si
‡ Ljubljana
skrjanec.iza@gmail.com

Povzetek

V prispevku predstavimo postopek, nabor oznak, merila ter orodje za semantično označevanje učnega korpusa za slovenščino. V prvem delu prispevka predstavimo teoretična izhodišča raziskave in uporabljeno metodologijo, nato pa podrobno opišemo nabor oznak za semantično označevanje učnega korpusa za slovenščino in merila za njihovo določanje. Posebej izpostavimo konkurenčne udeleženske vloge in potencialne nove udeleženske vloge za razreševanje mejnih primerov. Prispevek zaključimo s kratkim povzetkom sprejetih odločitev in predvidenim nadaljnjim delom v okviru bilateralnega projekta Označevanje semantičnih vlog v slovenščini in hrvaščini.

Semantic Role Labeling in the Training Corpus for Slovene

The paper describes the procedure, tagset, criteria and tools for semantic role labeling in the training corpus for Slovene. In the first part we present the theoretical foundations of our research and the methodology. The following part includes a detailed description of the tagset used for semantic role labeling of Slovene, together with annotation criteria. Ambiguous cases are highlighted and potential now semantic roles are suggested for solving borderline cases. The paper finishes with a short summary of the decisions that were taken in the process, and future work in the context of the bilateral Slovene-Croatian project Semantic Role Labeling in Slovene and Croatian.

1 Uvod

Označevanje semantičnih vlog (ang. Semantic Role Labeling – SRL) je postopek, ki je z jezikoslovnega vidika namenjen (avtomatskemu) prepoznavanju udeleženskih vlog, z jezikovnotehnološkega pa razvoju sistemov za luščenje informacij, sistemov za odgovarjanje na vprašanja (ang. question answering system), izboljšavi delovanja skladenjskih razčlenjevalnikov ter strojnih prevajalnikov ipd. (Shen in Lapata, 2007; Christensen et al., 2011). Ker pomanjkanje konsenza glede različnih kategorij in meril za njihovo določanje, ki so danes sicer že na voljo za številne jezike, povzroča težave pri čezjezikovnem modelu semantičnega označevanja, mora po našem mnenju uspešen sistem meril in oznak za označevanje udeleženskih vlog ali natančneje predikatno-argumentnih razmerij (a) zagotavljati nabor kategorij, ki je kar najbolj optimalen, tj. pokriti vse (v našem primeru za slovenščino) ključne udeleženske vloge in hkrati (b) ne vsebovati kategorij, ki so prepodrobne ali medsebojno prekrivne, (c) temeljiti primarno na semantičnih in ne na morfoloških, leksikalnih ali skladenjskih lastnostih, (d) omogočati formalni opis oz. uporabnost v jezikovnotehnoloških aplikacijah ter (e) biti čim bolj kompatibilen s kategorijami in merili, ki veljajo za druge jezike (prim. Petukhova in Bunt, 2008: 39). V ta namen je bil v okviru projekta izdelave učnega korpusa za označevanje semantičnih vlog za slovenščino izdelan sistem meril za prepoznavanje in označevanje udeleženskih vlog za slovenščino. Naš cilj je bil ročno označiti polovico skladenjsko označenega dela učnega korpusa ssj500k,[1] na njegovi podlagi pa naj bi bilo v prihodnje mogoče avtomatsko označiti tudi obsežnejše korpuse.

V nadaljevanju prispevka predstavimo izhodišča za določitev semantičnih kategorij ter nabor oznak za slovenščino, postopek označevanja in orodje za semantično označevanje učnega korpusa za slovenščino.

2 Teoretično in metodološko ozadje

Pri izbiri metode semantičnega označevanja in določanju semantičnih kategorij za slovenščino smo najprej analizirali posamezne pristope, ki so bili razviti in uporabljeni za druge jezike, npr. PropBank (Palmer et al., 2005), Verbnet (Kipper et al., 2006) in FrameNet (Backer et al., 1998) za angleščino, AnCora (Taulé et al., 2011) za španščino, SoNaR (Schuurman et al., 2010) za nizozemščino. Poleg tega pa še nabor oznak za hrvaščino (Filko et al., 2012) in češki valenčni leksikon Vallex.[2] Osredotočili smo se na primerjavo formalnih opisov (tj. naborov semantičnih oznak) za posamezne udeleženske vloge ter meril za njihovo določanje. Z vidika optimizacije nabora oznak, ki bi zagotavljal dovolj robusten sistem in hkrati v čim večji meri upošteval specifike slovenščine, smo upoštevali še stopnjo semantične razdrobljenosti, ki jo predvideva posamezni sistem, in dejstvo, da za slovenščino nimamo na voljo strojno berljivega leksikona glagolske vezljivosti. Poleg tega smo merila za semantično označevanje želeli določiti tako, da bodo

[1] Opis in prenos korpusa: http://www.slovenscina.eu/tehnologije/ucni-korpus.

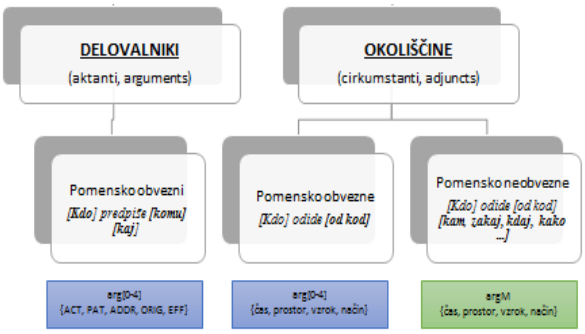
[2] Vallex: http://ufal.mff.cuni.cz/vallex.

omogočala zanesljivo in čim bolj konsistentno označenost učnega korpusa.

Pri naboru udeleženskih vlog in njihovih formalnih opisov smo izhajali iz funkcijskega generativnega pristopa Praške odvisnostne drevesnice (ang. Prague Dependency Treebank; PDT; Mikulová et al., 2006), ki znotraj obsega prototipičnega glagolskega vzorca (propozicije) določa razmerja med udeleženci, ki imajo lahko udeleženske vlogo delovalnika ali okoliščine. Udeležence in njihove različne udeleženske vloge določa torej prototipična prepozicija za dani pomen glagola, ki se uresničuje v njegovi vezljivostni shemi. Konkretno bi lahko rekli, da predvideva glagol *narediti* v enem od svojih pomenov tako delovalniške kot okoliščinske udeležence, ki se na oblikoskladenjski ravni realizirajo v obliki argumentov, ki jih je mogoče zapisati kot: kdo *naredi* komu kaj (kdaj, kje, kako, zakaj), kar predstavlja vezljivostno shemo konkretnega glagolskega pomena. Ob tem, da je število delovalnikov za dani primer predvidljivo, čeprav ne nujno realizirano v vsakem vzorcu omenjenega glagola oz. pomena, je mogoče za okoliščine reči le, da jih glagolski pomen predvideva, odprto pa ostaja vprašanje, ali so dejansko potrebne za ustrezno evokacijo glagolskega pomena. Na drugi strani je jasno, da tudi (ne)realizacija posameznih delovalnikov sama na sebi ne vpliva nujno na pomen glagola, saj govorec lahko udeležence predvideva, tudi če v vzorcu niso izraženi (Hanks, 2010; Žele, 2010). Tako velja, da obstaja v mentalnem leksikonu govorca za vsak glagolski pomen prototipična propozicija, ki se v realnem besedilu udejanja na različne načine, pri čemer izrablja tako oblikoskladenjski inventar jezika, vključno z elipsami in sobesedilnimi referencami, kot tudi zunajezikovne in pragmatične okoliščine izrekanja. Pri določanju razmerja med pomensko obveznimi in pomensko neobveznimi udeleženci smo zato v izhodišču izhajali iz sistema PropBank (Palmer et al., 2005), (Slika 1). Ta model določa pomensko obveznost le na ravni delovalnikov (določila), ki so vedno pomensko obvezni (arg[0-4]), medtem ko so okoliščine (prislovna dopolnila; argM) določene kot pomensko neobvezne, pri čemer, kot rečeno, pomenska obveznost ne implicira tudi strukturne obveznosti.

Slika 1: Obligatornost udeležencev v sistemih PropBank in VerbNet.

Na drugi strani sistemi, kot so PDT ter FrameNet (Backer et al., 1998), predvidevajo ugotavljanje obligatornosti tudi na ravni okoliščin (Slika 2). V tem primeru se tako pomensko obvezni delovalniki (ACT, PAT, ADDR, ORIG, EFF) kot pomensko obvezne okoliščine (čas, prostor, vzrok, način) označujejo z oznakami arg[0-4], pomensko neobvezne okoliščine pa z oznakami argM (gl. Tabelo 1).



Slika 2: Obligatornost udeležencev v sistemu PDT.

Ob združitvi obeh sistemov z vidika obligatornosti udeležencev, bi zgornji primer za glagol *narediti* izgledal takole:

Kdo	*naredi*	kaj	komu	kdaj	kje	kako	zakaj
arg0		arg1	arg2	{arg3-4}	{arg3-4}	{arg3-4}	
				ArgM			

Tabela 1: Vezljivostna shema glagola *narediti* s pripisom udeleženskih vlog in njihove obligatornosti v sistemu PDT.

Čeprav se zdi glede na pomenske lastnosti glagola ustreznejši sistem PDT, ki ob združitvi pomensko-skladenjskih meril prepoznava tudi obligatornost okoliščinskih udeležencev (prim. tudi Žele, 2010), obligatornosti za slovenščino ni bilo mogoče dosledno izpeljati brez leksikonskih podatkov za posamezni glagol. V trenutni fazi semantičnega označevanja smo zato ohranjali obligatornost le pri delovalnikih, pri okoliščinah pa te razmejitve nismo upoštevali, je pa to eden od izzivov, ki se jih nameravamo lotiti v nadaljnjih fazah označevanja.

3 Semantično označevanje za slovenščino: nabor oznak in merila za njihovo določitev

Osnovo za nabor udeleženskih vlog in njihovih oznak nam je, kot rečeno, predstavljal nabor oznak praške odvisnostne drevesnice. Z vidika optimizacije pomenske razdrobljenosti, upoštevanja slovenskih specifik in hkrati prekrivnosti oznak med posameznimi sistemi, smo nabor ustrezno zreducirali. Tabela 2 prikazuje združitev nabora delovalniških vlog glede na PDT ter nabor oznak za slovenščino.

Oznaka	PDT		SLO	
ACT	ACT	actor		**vršilec, aktant**
PAT	PAT	patient		**prizadeto**
REC	ADDR	addressee	prejemnik	**prejemnik**
	BEN	benefactor	koristnik	
ORIG	ORIG	origo	izvor	**izvor**
	HER	inheritence	podedovanost	
RESLT	EFF	effect		**učinek**

Tabela 2: Nabor oznak za delovalniške udeleženske vloge za slovenščino glede na PDT.

Praška odvisnostna drevesnica predvideva med okoliščinskimi udeleženci naslednje kategorije: čas, prostor, vzročnost in način, ki smo jih upoštevali tudi v

slovenskem naboru (Tabela 2). Pri notranji razčlenjenosti smo težili k združevanju pomenskih kategorij pod eno oznako. Tako smo npr. za različne časovne kategorije (*when, parallel, from when, to when; how long, since when* ipd. – skupaj 9), ki imajo v PDT ločene oznake, v slovenskem naboru združili v le tri: TIME (čas), DUR (trajanje) in FREQ (pogostnost). Hkrati smo določili, da oznaka TIME zajema semantične povezave, ki ustrezajo opredelitvam kot: *kdaj, sočasnost, z/od kdaj, do kdaj*, oznaka DUR določa povezave, ki opredeljujejo trajanje stanja ali dejanja (*kako dolgo, koliko časa*), oznaka FREQ pa pogostnost (*kako pogosto, kolikokrat*).

Tabela 3 prikazuje razmerje med PDT in slovenskimi oznakami za druge okoliščinske kategorije: prostor, vzrok in način.

	Oznaka	PDT		SLO	
PROSTOR	**LOC**	LOC	locative		**kraj**
		DIR2	which way	smer	
	SOURCE	DIR1	from		**začetna lokacija**
	GOAL	DIR3	where to		**končna lokacija**
VZROČNOST	**AIM**	AIM	aim	namen	**namen**
		INTT	intent	namera	
	CAUSE	CAUS	cause		**vzrok**
	CONTR	CNCS	concession	dopustnost	**protivnost**
		CONTRD	contradiction	protivnost	
	COND	COND	condition		**pogojnost**
OZIR	**REG**	REG	regard	ozir	**ozir**
		CRIT	criterion	merilo	
		CPR	comparison	primerjava	
	ACMP	ACMP	accompaniment		**spremstvo**
NAČIN	**RESTR**	RESTR	restriction	omejitev	**omejitev**
	MANN	MANN	manner	način	**način**
		RESL	result	rezultat	
	MEANS	MEANS	means		**sredstvo**
KOLIČ	**QUANT**	DIFF	difference	razlika	**količina**
		EXT	extent	količina	

Tabela 3: Nabor oznak za okoliščinske udeleženske vloge za slovenščino glede na PDT.

Pri označevanju večbesednih enot smo od nabora PDT ohranili le oznako DPHR (ang. dependant part of phraseme), ki smo jo preimenovali v PHRAS. Z njo označujemo frazeološke zveze tipa: *iti na živce*$_{PHRAS}$*, zaviti v molk*$_{PHRAS}$ ipd. Na novo smo v slovenskem naboru dodali oznako za zložene povedke MWPRED (ang. multi-word predicate), ki smo jo uporabili za označevanje zvez nedoločnika in faznega glagola, npr. *začeti vpiti*$_{MWPRED}$, ter za zveze nedoločnika in modalnega glagola, npr. *bo uspelo prepričati*$_{MWPRED}$*, zmore brati*$_{MWPRED}$*, niso želeli prikrajšati*$_{MWPRED}$. Zvez glagola in povedkovega prilastka nismo označevali s posebno oznako (PDT za te zveze uporablja oznako COMPL – ang. predicative complement), pač pa z delovalniško oznako (navadno RESLT): *zdela se mi je nekoliko bolj vsakdanja*$_{RESLT}$. Za označevanje modalnih glagolskih zvez smo uvedli oznako MODAL, npr. *je treba*$_{MODAL}$*, bi bilo mogoče*$_{MODAL}$. Zvez glagola z oslabljenim pomenom in samostalnika oz. samostalniške zveze (V PDT oznaka CPHR – ang. nominal part of the complex predicate) v tej fazi ne ločujemo od polnopomenskih vlog istih glagolov. To pomeni, da ne vzpostavljamo razlike med *dati ime* (PAT), *ne imeti namena* (PAT) – glagoli z oslabljenim pomenom – in *dati gol* (PAT), *imeti prijatelja* (PAT) – kjer gre za zvezo glagola in predmetnega določila. Razlikovanje med pomensko oslabljenimi glagoli kot sestavnimi deli glagolske zveze, prim. še *imeti na voljo, imeti v spominu,* in glagoli kot podeljevalci udeleženske vloge, *imeti denar*, bo prišlo do izraza pri prepoznavanju večbesednih enot, za katerega je predviden samostojni nivo označevanja učnega korpusa. To označevanje trenutno poteka v okviru COST akcije PARSEME kot del skupne naloge za identifikacijo večbesednih enot v različnih jezikih. Projekt predvideva v prvi fazi določitev formata in meril za označevanje večbesednih enot, v nadaljevanju pa ročno označenost približno 11.400 enot učnega korpusa.

Skupaj z oznakami smo na podlagi PDT določali tudi merila za njihovo prepoznavanje. V Tabeli 4 so poleg nabora oznak ter slovenskih imen zanje navedeni še kratki opisi udeleženskih vlog.

Udeleženska vloga		Opis
DELOVALNIKI	ACT	delujoči udeleženci, povzročitelji ali nosilci dejanja
	PAT	prizadeti predmet dejanja
	REC	prejemnik, posredni udeleženec dejanja; nedelovalniški udeleženec, ki mu je dejanje v škodo ali v prid
	ORIG	izhodišče, izvor/vir/povod dejanja
	RESLT	učinek, rezultat, cilj dejanja
OKOLIŠČINE	TIME	konkretni trenutek ali interval dejanja; kdaj
	DUR	trajanje stanja, dejanja koliko časa
	FREQ	frekvenca dejanja
	LOC	konkretna lokacija, kraj, mesto dejanja; kje
	SOURCE	začetna točka v prostoru; od kod
	GOAL	končna točka v prostoru; kam
	AIM	namen dejanja; čemu, s kakšnim namenom
	CAUSE	vzrok dejanja; zakaj
	CONTR	nepričakovana posledičnost dejanja; kljub čemu
	COND	pogoj za obstoj dejanja ali dogodka
	REG	glede na, primerjava
	ACMP	predmet, oseba ali dogodek, ki spremlja dejanje ali druge udeležence
	RESTR	izjema, omejitev
	MANN	načinovna lastnost dejanja, rezultat ob koncu dejanja
	MEANS	sredstvo ali orodje za izvedbo dejanja
	QUANT	količina, razlika
GLAGOLSKE ZVEZE	MWPRED	zveze z nedoločniki
	MODAL	zveze biti + modalnega prislova/pridevnika
	PHRAS	pomensko neprozorne zveze

Tabela 4: Merila za določanje udeleženskih vlog v učnem korpusu za slovenščino.

V slovenskem naboru smo od skupno 34 oznak v PDT ohranili 22 oznak (+ 2 za glagolske zveze), hkrati pa je analiza pokazala, da v nekaterih mejnih primerih potrebujemo podrobnejše pomenske opredelitve. V nadaljevanju opišemo nekatera ključna pomenska razmerja, ki zahtevajo natančnejšo opredelitev semantičnih oznak, in sicer tako z vidika pomenskih kot formalnih (skladenjskih in morfoloških) meril, ter potencialne dodatne/nove udeleženske vloge.

3.1 Konkurenčne udeleženske vloge/pomenska razmerja

Posamezna razmerja med udeleženci si z vidika določitve ustrezne udeleženske vloge pogosto konkurirajo. Do konkurenčnih povezav prihaja tako znotraj delovalnikov, npr. med vršilcem in prizadetim: *Dogodek v Ankaranu (ACT) je bila dramatična nesreča (PAT)*, kot tudi znotraj okoliščinskih razmerij, npr. med prostorskimi (LOC) in vzročnimi (CAUSE) ali časovnimi (TIME) udeleženskimi vlogami: *se dušijo v številkah (LOC→CAUSE), ministrica je na enem od sestankov (TIME→LOC) dejala*, ter med delovalniškimi in okoliščinskimi udeleženci, npr. pri prostorsko izraženih delovalnikih: *v bolnišnici (LOC→ACT] bodo uvedli,* o čemer več v nadaljevanju.

Razmerje vršilec – prizadeto pride do izraza predvsem pri upoštevanju skladenjskih razmerij, npr. pri razlikovanju med pasivnimi in aktivnimi zgradbami, kjer ločujemo med dvema oblikoskladenjskima možnostima izražanja trpnosti, tj. s prostim glagolskim morfemom *-se/-si* in z deležnikom na *-n/-t,* ter med trpnimi in aktivnimi povratnosvojilnimi skladenjskimi zgradbami. Na podlagi tega obravnavamo primere kot npr. *pozitivna diskriminacija* (PAT) *se označuje kot privilegij* kot pasivne, primere kot *dogodki* (ACT) *so se odvijali* pa kot aktivne povratnosvojilne skladenjske zgradbe. Pri trpnih zgradbah smo pozorni na to, da ostajajo udeleženske vloge v aktivnih in pasivnih skladenjskih zgradbah prekrivne, npr. *stvar* (PAT) *je malce bolj zapletena – kdo* (ACT) *zaplete stvar* (PAT); *pozitivna diskriminacija* (PAT) *se označuje kot privilegij – kdo* (ACT) *označuje pozitivno diskriminacijo* (PAT) *kot privilegij.*

V pomenskih razmerjih med delovalniškimi in okoliščinskimi udeleženci si konkurirajo zlasti predložne delovalniško-prostorske pomenske povezave kot npr. LOC→ACT/REC/RESLT: *v bolnišnici* (LOC→ACT) *so uvedli; povzročiti nevšečnosti na televiziji* (LOC→REC); *zaposliti v podjetju* (LOC→RESLT), in GOAL→ RESLT/REC: *to ga je spravilo v dobro voljo* (GOAL→RESLT); *Nokia je v paket* (REC→GOAL) *priložila polnilnik*. O prostorsko izraženih aktantih govorimo takrat, ko glagolski pomen ne predvideva prostorske komponente, kot jo predvidevajo npr. glagoli premikanja *priti, oditi, iti* itd. KAM, ampak dejavnost, npr. *v osnovni šoli* (ACT) *pripravljajo, pri Fujifilmu* (ACT) *so objavili, na centru* (ACT) *so se lotili*. Aktanti so v teh primerih dejansko metonimično izraženi delovalniki, hkrati pa vezljivostni vzorec glagola predvideva tudi potencialno okoliščinsko udeleženske mesto, ki se v nekaterih primerih tudi dejansko izraža, npr. *v osnovni šoli v Bistrici* (ACT) *pripravljajo.*

3.2 »Manjkajoče« udeleženske vloge/pomenska razmerja

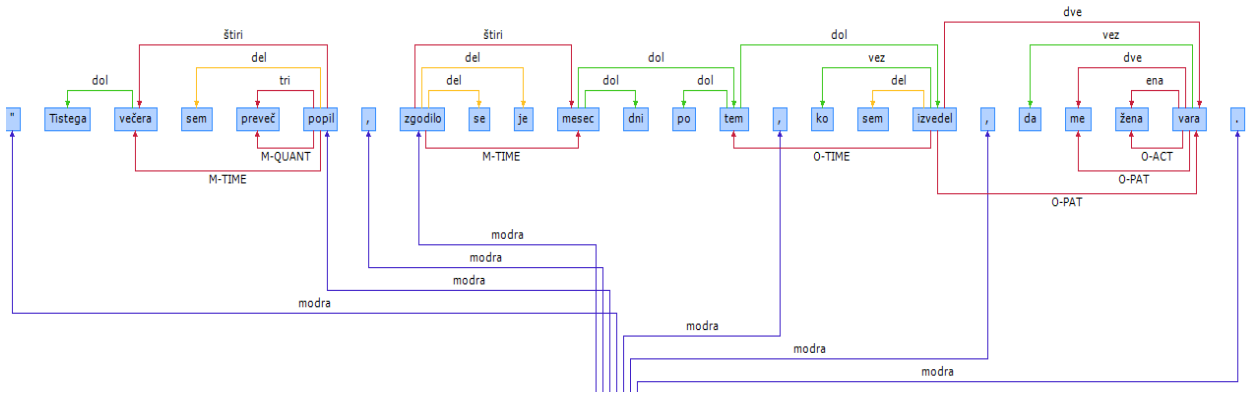
Pri glagolskih pomenih, ki predvidevajo obsežnejši vezljivostni vzorec, kot so npr. glagoli sporočanja, zaznavanja, mišljenja ipd., npr. kdo-ACT *reče, pove, izjavi* ... komu-REC kaj-RESLT o čem-PAT (kdaj, kje, kako, zakaj), se glede na možnost pripisa iste udeleženske vloge samo enemu udeležencu pojavlja potreba po pomensko podrobnejši razčlenitvi udeleženskih vlog.

Sem sodi razlikovanje med konkurenčnimi okoliščinskimi udeleženskimi vlogami, ki smo jih omenjali že v prejšnjem poglavju. V stavkih kot: *v drevišnjih tekmah bodo igrali*; *sploh ni padel v vojni*; *na enem od sestankov je dejala*, se podčrtanim udeležencem pripisuje bodisi časovna (TIME) bodisi prostorska (LOC) udeleženska vloga. Ker ob tem vezljivostni vzorec lahko predvideva več časovnih ali prostorskih udeležencev, kot npr. v stavku *Na veleslalomu za mladince na SP na Pohorju*, se odpira vsaj še prazno mesto »dogodka« (tekma, vojna, sestanek, SP), kot ga denimo predvideva sistem FrameNet (EVENT), ki združuje tako prostorsko kot časovno pomensko komponento. Podobno se za konkurenčni udeleženski vlogi načina (MANN) in prostora (LOC) v primerih kot: *informacijo po elektrodnem kablu pripeljejo v napravo; mimo se je pripeljala deklica,* ponuja možnost bolj specializirane udeleženske vloge, npr. »pot« (PATH), ki jo prav tako pozna sistem FrameNet.

4 Orodje in format označevanja učnega korpusa za slovenščino

Za semantično označevanje korpusa smo uporabili orodje SentenceMarkup, ki je bilo primarno razvito za skladenjsko označevanje slovenščine (Dobrovoljc et al., 2012). Orodje smo prilagodili za namene semantičnega označevanja tako, da smo mu dodali neodvisen in hkrati medsebojno povezljiv semantični nivo (Slika 3).

Ker želimo program v prihodnje nadgraditi za različne tipe označevanja (npr., za označevanje večbesednih enot), je pomembno, da zagotavlja čim večjo avtonomnost pri spreminjanju nabora oznak na več ravneh in možnosti tako ločenega kot kombiniranega iskanja po tipih povezav na skladenjski, pomenski ter drugih ravneh označevanja. Program omogoča izvoz podatkov v tabelarični obliki in XML formatu, ki poleg podatkov o tipu povezave na posameznem nivoju označevanja vsebuje tudi podatke o lemi, MSD-oznake ter omogoča izpis celotnega stavka.



Slika 3: Skladenjski in semantični nivo oznak v orodju SentenceMarkup.

5 Zaključek in prihodnje delo

V trenutni fazi semantičnega označevanja učnega korpusa je bil naš cilj določiti dovolj robusten in hkrati optimalen nabor udeleženskih vlog za slovenščino. Nabor oznak in merila za njihovo označevane smo določili na podlagi obstoječih označevalskih modelov, kjer smo izhajali zlasti iz PDT, v posameznih odločitvah pa smo upoštevali tudi rešitve v sistemu FameNet in drugih.

V postopku ročnega označevanja učnega korpusa smo, izhajajoč iz dejstva, da ne razpolagamo z leksikonom glagolske vezljivosti za slovenščino, težili k izboru udeleženskih vlog, ki omogočajo konsistentnost označevanja z upoštevanjem tako skladenjskega kot pomenskega nivoja. Pri konkurenčnih pomenskih oznakah smo zato skušali uvesti čim bolj jasna razločevalna merila, hkrati pa smo predlagali dodatne udeleženske oznake, ki razrešujejo mejne primere. V prihodnje je naš namen na podlagi analize semantičnih povezav določiti tudi stopnjo obligatornosti tako pri delovalniških kot tudi pri okoliščinskih udeleženskih vlogah.

V naslednji fazi nameravamo v okviru bilateralnega projekta med Slovenijo in Hrvaško oblikovati sistem za označevanje semantičnih vlog, ki bodo pripisane obstoječim skladenjskim odvisnostnim povezavam v učnih korpusih, ki so uporabljeni za algoritme strojnega učenja za oba jezika. Vzorčni del slovenskega in hrvaškega učnega korpusa bo označen s kompatibilnimi oznakami, na njih pa bodo izpeljani tudi prvi eksperimenti avtomatskega označevanja z nadzorovanim strojnim učenjem. V okviru projekta bodo tako izdelana skupna navodila za označevanje semantičnih vlog v slovenščini in hrvaščini, orodje za označevanje semantičnih vlog za označevalce na obeh straneh, vzorčna učna korpusa za slovenščino in hrvaščino in eksperimentalno orodje, ki uporablja strojno učenje, za avtomatizacijo označevanja semantičnih vlog.

Del korpusa ssj500k je bil novembra 2015 vključen v zbirko skladenjskih drevesnic Universal Dependencies (UD) (Nivre et al., 2016). To omogoča, da se sistem označevanja semantičnih vlog, razvit v obstoječem sistemu JOS, prenese in preveri tudi v sistemu UD, kar je ena od prihodnjih nalog. Prenos je smiseln tudi s stališča kompatibilnosti med slovenskim in hrvaškim sistemom označevanja v okviru bilateralnega projekta, saj hrvaška drevesnica uporablja oznake UD.

6 Literatura

Collin F. Backer, Charles J. Fillmore in John B. Lowe. 1998. The Berkeley FrameNet project. *Proceedings of the COLING-ACL*. Montreal, Canada. 86–90.

Janara Christensen, Stephen Soderland Mausam in Oren Etzioni. 2011. An Analysis of Open Information Extraction based on Semantic Role Labeling. *International Conference on Knowledge Capture* (KCAP). Banff, Alberta, Canada. June 2011. 113–120.

Kaja Dobrovoljc, Simon Krek in Jan Rupnik. 2012. Skladenjski razčlenjevalnik za slovenščino. V T. Erjavec, J. Žganec Gros (ur.): *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 42–47.

Matea Filko, Daša Farkaš in Danijela Merkler. 2012. *SRL Tagset for Croatian*. Institute of Linguistics, Faculty of Humanities and Social Sciences, Zagreb. http://hobs.ffzg.hr/static/docs/SRL_tagset.pdf.

Patrick Hanks. 2010. Elliptical Arguments: a Problem in relating Meaning to Use. S. Granger, M. Paquot (ur.): eLexicography in the 21st century: New challenges, new applications. *Proceedings of ELEX2009*. Cahiers du CENTAL. Louvain-la-Neuve: Presses universitaires de Louvain.

Karin Kipper, Anna Korhonen, Neville Ryant in Martha Palmer. 2006. Extensive Classifications of English verbs. *Proceedings of the 12th EURALEX International Congress*. Turin, Italy. September. 1–15.

Marie Mikulová et al. 2006. *Annotation on the tectogrammatical level in the Prague Dependency Treebank*. Aannotation manual. Technical Report 30. 5–11.

Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. V: *Proceedings of LREC'16*. 1659–1666.

Martha Palmer, Daniel Gildea in Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics* 31(1). 71–106.

Volga Petukhova in Henry Bunt. 2008. LIRICS semantic role annotation: Design and evaluation of a set of data categories. V *Proceedings of the 6th International Conference on Language Resources and Evaluation* (LREC 2008). Marrakech, Morocco. Paris: ELRA. 39–45.

Ineke Schuurman, Véronique Hoste in Paola Monachesi. 2010. Interacting semantic layers of annotation in sonar, a reference corpus of contemporary written dutch. *Proceedings of LREC'10*, Valletta, Malta. ELRA. 2471–2477.

Dan Shen in Mirella Lapata. 2007. Using Semantic Roles to Improve Question Answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning*. Prague. 12–21.

Mariona Taulé, Antònia M. Martí in Oriol Borrega. 2011. AnCora 2.0: Argument Structure Guidelines for Catalan and Spanish. *Working paper 4: TEXT-MESS 2.0 (Text-Knowledge 2.0)*. Universitat de Barcelona. Barcelona.

Andreja Žele. 2010. Elipsa med glagolsko intenco in besedilno koherenco (Izpust med glagolsko usmerjenostjo in besedilno soveznostjo). *Slavistična revija*, 58(1). 117–131.