SEMANTIČNO OZNAČEVANJE KORPUSA SLOVENŠČINE PO MODELU FrameNet

Sara Može
Koper

UDK 811.163.6'322:811.163.6'37

FrameNet je obsežen računalniški leksikografski projekt, ki temelji na teoriji pomenskih shem in katerega cilj je podati popoln opis povezanih skladenjskih in semantičnih lastnosti angleških leksikalnih enot. V pričujočem članku bo predstavljena pilotna študija o uporabi modela FrameNet pri semantičnem označevanju manjšega korpusa slovenščine. Prikazane bodo rešitve, ki nam jih za posamezne jezikovne probleme ponuja model, in predstavljene njegove prednosti in pomanjkljivosti.

FrameNet, korpus, pomenska shema, semantično označevanje, semantika

FrameNet is a large computational lexicography project based on Frame Semantics. Its goal is to provide a full description of the linked syntactic and semantic proprieties of English lexical units. The article presents a pilot study on the use of the FrameNet model in the semantic annotation of a small corpus of Slovene. Specific linguistic problems and solutions provided by the model will be presented, and both its advantages and disadvantages discussed.

FrameNet, corpus, frame, semantic annotation, semantics

1 Uvod

FrameNet je računalniški leksikografski projekt, ki ga od leta 1997 razvijajo na Mednarodnem inštitutu za računalniške vede v Berkeleyju. Cilj projekta je iz obsežnih besedilnih korpusov z ročnimi in avtomatskimi postopki pridobivati informacije o povezanosti skladenjskih in semantičnih lastnosti angleških leksikalnih enot. Končni rezultat je podatkovna baza, ki obsega 10.000 leksikalnih enot, od katerih jih je 6.100 v celoti označenih, in 135.000 označenih povedi, vzetih pretežno iz korpusa The British National Corpus in delno iz korpusov LDC North American Newswire. Bazo FrameNet uporablja na stotine raziskovalcev, učiteljev in študentov po vsem svetu, v okviru raziskovalnih projektov pa je nastalo že več podatkovnih baz za druge jezike, npr. za španščino, nemščino in japonščino.

2 Zasnova projekta

Teoretično izhodišče projekta je *teorija pomenskih shem* (v ang. Frame Semantics), ki jo je utemeljil znani ameriški jezikoslovec Charles J. Fillmore. Teorija sloni na prepričanju, da ne moremo poznati pomena besede ali besedne zveze, če ne poznamo tudi celotne zunajjezikovne situacije, ki je nanjo vezana. Namesto da bi o pomenu besede ali besedne zveze sklepali na podlagi jezika, torej tako, da opazujemo, kako se beseda oziroma besedna zveza »obnaša« v sobesedilu, najprej oblikujemo opis situacije in mu šele nato pripišemo ustrezno besedo ali besedno zvezo. Shematično predstavitev dane situacije imenujemo *pomenska shema* (v ang.

frame),[1] udeležence in okoliščine v opisani situaciji pa *shemski elementi* (v ang. frame elements ali na kratko FE).[2] Znotraj teorije ne govorimo o paru beseda – shema, temveč pomen – shema. Pri večpomenskih besedah je denimo vsak pomen posebej vezan na drugačno zunajjezikovno situacijo, zato je logično, da mu bo »pripadala« druga shema. Če bi denimo imeli opravka z besedo *bister*, bi pomena *nemoten* in *pameten* pripadala dvema različnima shemama. Za leksikalno enoto, torej besedo ali besedno zvezo (frazni glagol, idiom ali katero drugo večbesedno enoto) v točno določenem pomenu, pravimo, da *prikliče* pripadajočo shemo; beseda *bister* lahko torej prikliče več shem, katero od shem pa bo v posameznem primeru priklicala, je odvisno od sobesedila (Fillmore, Johnson, Petruck 2003: 235–236, Petruck 1996, Ruppenhofer idr. 2006: 5–9).

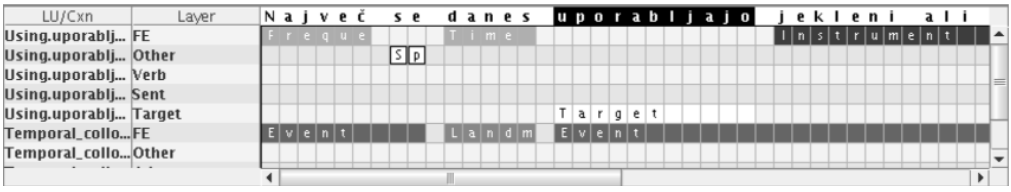
Za primer vzemimo situacijo, v kateri darovalec prenese nek predmet na obdarovanca, tako da predmet preide iz njegove v obdarovančevo last. Znotraj FrameNeta situacijo predstavlja shema Giving (darovanje), posameznim udeležencem v situaciji pa pripadajo shemski elementi DAROVALEC (v angleščini mu pravimo DONOR), PREDMET (v angleščini THEME) in OBDAROVANEC (v angleščini RECIPIENT). Shemo prikličejo leksikalne enote, kot so glagoli *dati*, *darovati* in *izročiti* in samostalniki *darilo*, *darovalec* in *donacija*, npr.:

[1] In zdi se, da ji je rekel mornar:
»Pridi na barko,
[zdravilo za tvojega otročička $_{\text{PREDMET}}$]
[ti $_{\text{OBDAROVANEC}}$] bom dal!«[3]

Sistem, ki ga za semantični opis ponuja FrameNet, je osnovan na človeškem izkustvu; ker ni jezikovno specifičen, ga lahko uporabimo za različne jezike, torej tudi za slovenščino.

3 Označevanje v FrameNetu

Znotraj FrameNeta se besede označuje tako, da se v vzporedno poravnanih plasteh dodajajo tipološko različne oznake. Ker ustvarjalce sistema v prvi vrsti zanima razmerje med skladnjo in semantiko, so morali poleg plasti, v kateri se določajo shemski elementi, razviti še skladenjske oznake dveh vrst, in sicer so to oznake za besedne zveze (plast PT – iz ang. phrase type) in skladenjske funkcije besednih zvez v povedi (plast GF – iz ang. grammatical function). To so osnovne plasti označevanja. Poleg njih obstajajo v sistemu še dodatne plasti, v katerih označujemo posebne jezikovne pojave, npr. plast Other, v kateri označujemo prilastkove odvisnike in angleške eksistencialne povedi, in posebna plast, ki se pri označevanju samostalnikov imenuje Noun, glagolov Verb, pridevnikov Adj (Adjective), prislovov Adv (Adverb) in predlogov Prep (Preposition) (Ruppenhofer idr. 2003: 19–20). Označevanje



Slika 1 – Plasti in oznake med označevanjem v programu FrameNet Desktop.

1 Slovenska terminologija je povečini povzeta po Kreku (2008), ki je FrameNet tudi že predstavil slovenskemu bralstvu.
2 V slovnici globinskih sklonov (Fillmore 1968) se uporablja termin *globinski skloni*, v tvorbeno-pretvorbeni slovnici termin *udeleženske vloge* (Jackendoff 1990), Tesnière (1959) pa uporablja izraza *aktanti* in *cirkumstanti*. Vsi ti izrazi se nanašajo na semantično vezljivost oziroma na strukturo semantičnih argumentov.
3 Vsi slovenski primeri so povzeti po označenem gradivu.

v teh dveh plasteh bo podrobneje predstavljeno v nadaljevanju.

4 Semantično označevanje slovenščine

Model FrameNet je bil uporabljen pri semantičnem označevanju manjšega korpusa slovenščine. Korpus je nastal znotraj projektov Jezikoslovno označevanje slovenščine[4] in Sporazumevanje v slovenskem jeziku,[5] sestavljen je iz petsto naključno izbranih povedi iz besedilnozvrstno raznolikega nabora slovenskih besedil in je že pred semantičnim označevanjem vseboval oblikoskladenjske in skladenjske oznake,[6] zaradi česar mu oznak za tipe besednih zvez (PT) in skladenjske funkcije (GF) ni bilo treba dodajati. Označevanje je potekalo v programu FrameNet Desktop, ki so ga razvili na Mednarodnem inštitutu za računalniške vede v Berkeleyju. Končni rezultat je v celoti označenih dvesto povedi iz korpusa oziroma skupno 908 leksikalnih enot. V nadaljevanju bo predstavljenih nekaj zanimivejših jezikovnih problemov, ki so se pojavili pri označevanju.

4.1 Pomensko izpraznjeni glagoli

Znotraj FrameNeta so razvili nabor oznak, ki omogoča celovito označevanje struktur s pomensko izpraznjenimi besedami. Poseben tip takšne strukture so zveze pomensko izpraznjenih glagolov in pomensko bogatih samostalnikov, npr. *dati izjavo*, v katerih je semantično jedro zveze samostalnik in ne glagol. Zveze lahko pogosto parafraziramo z enim (najpogosteje izsamostalniškim) glagolom, npr. *dati izjavo* – *izjaviti*, *biti mnenja* – *meniti*, *imeti izkušnje* – izkusiti, *valiti krivdo* – *okriviti*, *voditi na sprehod* – *sprehajati* itn. Za tovrstne zveze je značilno, da samostalnik že sam po sebi izraža nek dogodek, neko stanje ali razmerje in je skladenjsko vezan na dani glagol, glagol pa je pomensko izpraznjen, kar pomeni, da pomen zveze izvira skoraj povsem iz samostalnika. Funkcija glagola je ta, da samostalnik skladenjsko *podpira*, zato mu znotraj FrameNeta pravimo *podporni glagol* (v ang. support verb).

Pomensko gledano lahko pomensko prozorne oziroma podporne glagole razdelimo na:

– navadne (glagol v zvezo vnaša zelo malo semantike, npr. *biti mnenja*, *imeti izkušnje*, *početi grdobije*),

– vidske oziroma fazne (glagol določa fazo nekega dejanja, dogajanja ali stanja, npr. *ostati umirjen*, *ostati vlažen*,[7] *postati par*, *postaviti nakupne naloge*, *uvajati pojem*, *zbrati pogum*),

– glagole, ki izražajo nek zorni kot (npr. *valiti krivdo* v prim. s *sprejeti krivdo*),

– slogovno zaznamovane (npr. *voditi na sprehod* v prim. s *peljati na sprehod*),

– povzročilne (npr. *have a headache* (imeti glavobol) v prim. z *give a headache* (dati nekomu glavobol oziroma nekomu povzročiti glavobol) (Ruppenhofer idr. 2006: 53–54).

V FrameNetu tovrstne glagole označujemo tako, da jim pri označevanju pripadajočega samostalnika v plasti Noun dodamo oznako Supp (ali Support; v prevodu podporni glagol), pri označevanju glagola pa ga kot leksikalno enoto zaznamujemo s semantičnim tipom Support. Semantični tip (v ang. semantic type) je posebna vrsta oznake, ki opozarja na posebne semantične lastnosti izbrane leksikalne enote; s tem ko glagol *valiti* označimo s semantičnim tipom Support, sugeriramo, da se lahko pojavlja v funkciji podpornega glagola.

4 Spletna stran projekta: <http://nl.ijs.si/jos.>.
5 Spletna stran projekta: <http://www.slovenscina.eu.>.
6 Oba tipa oznak sta bila razvita znotraj omenjenih projektov. Za več informacij gl. *Oblikoskladenjske specifikacije* JOS: <http://nl.ijs.si/jos/josMSD-sl.html.>.
7 Označevanje pomensko izpraznjenih glagolov ob pridevnikih v strokovni literaturi sicer ni omenjeno, vendar pa je iz podatkov razvidno, da so jih označevali na enak način kot glagole ob samostalnikih.

4.2 Nadzorni glagoli

Nadzorni glagoli (v ang. controllers) so posebna vrsta podpornih glagolov, ki se pojavljajo ob samostalnikih. Od glagolov, ki jih označujemo z oznako Supp, se razlikujejo v tem, da prikličejo drugačno shemo kot samostalnik in si z njim delijo enega od udeležencev, npr.:

[2] Za ta pogled bi bilo čudno, če bi bilo **nadzor** sploh kdaj mogoče uspešno [*izvrševati* $_{\text{CTRLR}}$] in [*ohraniti* $_{\text{CTRLR}}$].

V zgornjem primeru sta glagola *izvrševati* in *ohraniti* nadzorna glagola samostalnika *nadzor*. Če ju primerjamo z glagolom *zbrati* v zvezi *zbrati pogum*, kjer je bil označen s Supp, je razlika očitna: zveza *zbrati pogum* opisuje le eno situacijo, medtem ko zvezi *izvrševati nadzor* in *ohraniti nadzor* opisujeta dve – nekdo izvršuje nekaj in hkrati tudi nadzoruje oziroma nekdo ohranja nekaj in hkrati nekaj nadzoruje. Znotraj sistema jim v plasti Noun pripišemo oznako Ctrlr (ali Controller).

4.3 Neskladja med strukturo semantičnih in skladenjskih argumentov

Pri nekaterih zapletenejših samostalniških besednih zvezah skladenjsko jedro zveze ne sovpada s semantičnim – natančneje gre za zveze tipa *vrsta ukrepov*, v kateri je prvi samostalnik skladenjsko jedro zveze in je pomensko izpraznjen, saj zgolj določa neko lastnost drugega samostalnika, ki je semantično jedro zveze. V *Slovenski slovnici* so skladenjsko opredeljeni kot zveze samostalnikov in desnega neujemalnega sklonskega nepredložnega prilastka v rodilniku (Toporišič 1991: 468). V slovenščini lahko samostalniki, ob katerih se ti prilastki pojavljajo, med drugim izražajo količino, npr. *množina*, *ducat*, *vrsta*, *število*, *del*, *večina* itn., mero, npr. *kila*, *meter*, *vreča*, *kozarec*, *pest*, *žlica*, *skleda*, *kos* itn., in vrsto, npr. *vrsta*, *rod*, *pleme*, *sorta*, *tip* itn. (Toporišič 1991: 249–250).[8] Ker je razlikovanje med skladenjskim in semantičnim jedrom tovrstnih zvez bistveno pri določanju shemskih elementov, je prvi samostalnik označen s tipom Transparent_noun (izraz pomeni *pomensko izpraznjeni samostalnik*) (Ruppenhofer idr. 2006: 114–115). Oglejmo si primer:

[3] Oblasti v Šanghaju, enem največjih mest na Kitajskem, so se zelo resno zavzele za izboljšanje podobe mesta in v ta namen **sprejele** celo [vrsto $_{\text{Transparent\_noun}}$] ukrepov.

Jasno je, da oblasti niso sprejele *vrst*, temveč *ukrepe*; s tem ko prvi samostalnik *vrsto* s semantičnim tipom označimo za pomensko izpraznjenega, sugeriramo, da je semantično jedro zveze drugi samostalnik in ga na ta način povežemo z glagolom (*sprejeti – ukrep*).

4.4 Razlike med jezikovnima sistemoma slovenščine in angleščine

Pri označevanju se je za problematično že na začetku izkazalo označevanje prostih morfemov *si* in *se* v večdelnih glagolih. Angleščina tovrstnih prostih morfemov ne pozna, zato FrameNet zanje ne predvideva posebnih oznak. V korpusu so bili glagoli s prostima morfemoma označeni na tri načine. Glagoli, v katerih je prosti morfem neizpustljiv, npr. *pojaviti se*, so bili skupaj s prostim morfemom obravnavani kot ena večbesedna leksikalna enota. V to skupino spadajo tudi glagoli tipa *delati se* (v smislu *pretvarjati se*), v katerih prosti morfem bistveno spremeni pomen glagola in je torej dejansko neizpustljiv. Prosti morfemi v glagolih tipa *ubiti se*, v katerih prosti morfem dela prvotno prehodni glagol neprehoden, in *umivati se*, v katerih je prosti morfem mogoče zamenjati z zaimkom, npr. *sebe*, ali samostalnikom (sem spadajo tudi recipročni glagoli),[9] so bili označeni kot shemski elementi glagola, npr.:

8 V *Slovenski slovnici* so navedene še tri druge skupine samostalnikov, ki pa pomensko niso izpraznjeni in ne opisujejo samostalnika v prilastku.

9 Kategoriji sta povzeti po *Slovenski slovnici* (Toporišič 1991: 294–295).

[4] Šest let je že, odkar sva [se $_{\text{INDIVIDUALS}}$] **spoznala** na fitnesu na Jesenicah.[10]

Za najtežavnejše so se izkazali glagoli, v katerih prosti morfem izraža splošnega vršilca dejanja, npr.:

[5] Največ [*se* $_{\text{SPECIAL-OTHER}}$] danes **uporabljajo** jekleni ali aluminijasti okvirji različnih profilov.

Za jezikovna sredstva, ki izražajo splošnega vršilca dejanja, FrameNet ne predvideva posebnih oznak; ker je tovrstno rabo prostih morfemov smiselno zabeležiti, je bila morfemom pripisana oznaka Special-Other (posebno-drugo) v plasti Other, vendar pa bi bilo treba v prihodnosti zanje razviti novo oznako.

5 Sklep

Najpomembnejša prednost modela FrameNet je njegova univerzalnost, saj lahko semantično raven opisa učinkovito uporabimo za kateri koli jezik. Po drugi strani je treba poudariti, da je nabor skladenjskih oznak, ki v pričujočem članku sicer ni bil predstavljen (za več informacij o skladenjskih oznakah gl. Krek 2008), specifičen za angleščino in bi ga bilo treba za slovenščino korenito spremeniti, prav tako pa bi bilo treba dodati oznake za nekatere posebne jezikovne pojave, kot sta denimo prosta morfema *si* in *se*. Kljub vsemu FrameNet omogoča izčrpen semantični opis leksikalnih enot, končni izdelek, tj. podatkovna baza, pa je uporaben tako za navadne uporabnike kot za razvijalce jezikovnotehnoloških računalniških sistemov in jezikoslovce na številnih področjih, tj. od slovničarjev, leksikologov in leksikografov do prevajalcev, ki si z njim pomagajo pri alternativnih prevodnih ubeseditvah.[11] Ena od pomanjkljivosti je ta, da je označevanje zamudno in izredno zahtevno, v samem sistemu pa je bilo zaslediti tudi nekaj napak in nepopolnosti, npr. pomanjkljive opise nekaterih shem in podvajanje drugih, poleg vsega pa bi bilo treba sistem dopolniti z novimi shemami, saj določenih leksikalnih enot ni bilo mogoče označiti, ker shema zanje še ne obstaja. Kljub pomanjkljivostim lahko model ocenimo za več kot zadovoljiv; FrameNet predstavlja velik potencial za slovenščino in bi ga bilo kot dragocen vir jezikovnih podatkov vredno razvijati tudi v prihodnje.

Literatura

FILLMORE, Charles J., 1968: The Case for Case. Bach, Emmon, Robert T. Harms (ur.): *Universals in Linguistic Theory*. New York: Holt, Rinehart and Winston. 1–88.

FILLMORE, Charles J., JOHNSON, Christopher R., PETRUCK, Miriam R. L., 2003: Background to FrameNet. Fontenelle, Thierry (ur.): *FrameNet and Frame Semantics. International Journal of Lexicography*. Oxford: Oxford University Press. 235–250.

JACKENDOFF, Ray, 1990: *Semantic Structures*. Cambridge: MIT Press.

KREK, Simon, 2008: FrameNet in slovenščina. *Jezik in slovstvo* 53/5. 37–54.

PETRUCK, Miriam R. L., 1996: Frame Semantics. Verschueren, Jef, idr. (ur.): *Handbook of Pragmatics*. Philadelphia: John Benjamins.

Projekt FrameNet: <http://framenet.icsi.berkeley.edu.>. (Dostop 9. 7. 2009.)

RUPPENHOFER, Josef, idr., 2006: *FrameNet II: Extended Theory and Practice*: <http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126.>. (Dostop 9. 7. 2009.)

TESNIÈRE, Lucien, 1959: *Elements de syntaxe structurale*. Paris: Klincksieck.

TOPORIŠIČ, Jože, 1991: *Slovenska slovnica*. Maribor: Obzorja.

10 Morfem je označen kot shemski element OSEBE (v angleščini INDIVIDUALS) v shemi Make_acquaintance (spoznati (nekoga)).

11 Prevajalci lahko bazo uporabljajo kot orodje za parafraziranje, tako da danemu izrazu poiščejo bodisi sopomenko, protipomenko in nadpomenko bodisi drugo strukturo, npr. zvezo pomensko izpraznjenega glagola in pomensko bogatega samostalnika namesto polnopomenskega glagola (gl. točko 4.1).