10.35539/LTNC.2022.0047.B.S.G.N.B.L.R.V.J.L.N.G.284.287

Latviešu valodas *FrameNet* korpuss

Baiba Saulīte, Gunta Nešpore-Bērzkalne, Laura Rituma, Viesturs Jūlijs Lasmanis, Normunds Grūzītis

Valsts pētījumu programmas "Humanitāro zinātņu digitālie resursi" projekts Nr. VPP-IZM-DH-2020/1-0001 "Humanitāro zinātņu digitālie resursi: integrācija un attīstība". Darbs uzsākts ERAF Pēcdoktorantūras pētniecības atbalsta projektā Nr. 1.1.1.2/VIAA/1/16/188 "No abstraktās nozīmes reprezentācijas līdz dabiskam teikumam un saistītam tekstam" un ERAF praktiskas ievirzes pētījumu programmas projektā Nr. 1.1.1.1/16/A/219 "Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā".

Atslēgvārdi: semantiskā marķēšana, freimu semantika, semantiskā loma, leksiskā vienība, verbs, deverbāls atvasinājums

FrameNet ir starptautiski atzīta un dažādām valodām izmantojama valodas datu marķēšanas sistēma, kuras pamatā ir freimu semantika – leksiskās nozīmes teorija, kas veidota uz Čārlza Filmora (*Charles J. Fillmore*) un viņa kolēģu darbu pamata (Fillmore 1976; Fillmore et al. 2003; Fillmore and Baker 2010). Atbilstoši *FrameNet* (turpmāk tekstā – FN) pieejai tiek uzskatīts, ka katra vārda nozīmi vislabāk var saprast semantiskā situācijā jeb freimā – vispārinātā notikuma vai attieksmju modelī, kurā aprakstīti iespējamie situācijas dalībnieki un tādi situācijas apstākļi kā laiks, vieta u. c.

Kopš 2017. gada tiek veidots Latviešu valodas *FrameNet* korpuss (turpmāk tekstā – FrameNet-LV), tajā manuāli marķētas FN sistēmā definētās semantiskās lomas, vienlaikus sastatot biežāk lietotos latviešu valodas verbus (1350 leksēmu) un deverbālos lietvārdus (250 leksēmu) ar tiem atbilstošajiem angļu valodas FN freimiem.

Šajā rakstā aplūkota freimu semantikas pieeja datu marķēšanā, aprakstīta FrameNet-LV izveide – datu atlase, marķēšanas principi, kā arī parādīts, kā datus var analizēt, izmantojot korpusa tīmekļvietni.

Freimu semantika

Atbilstoši FN pieejai teksta korpusā katrai izvēlētajai leksiskajai vienībai (nozīmei) tiek norādīts tai atbilstošais freims, kā arī tiek marķēti tā freima elementi (turpmāk tekstā – FE) jeb semantiskās lomas. Piemēram, cilvēka dzimšanas situāciju (BEING BORN ‘piedzimt’) raksturo freims, kurā ir FE Bērns (cilvēks, kurš piedzimis), Laiks un Vieta (situācijas apstākļi). Iespējamas arī sarežģītākas situācijas, kurās ir vairāk FE, piemēram, darba uzsākšanas situācija (HIRING ‘pieņemšana darbā’), kurā ir šādi FE: Darbinieks, Darba devējs, Darba vieta, Amats, Pienākums, Laiks, Vieta, Veids utt.

Atkarībā no tā, vai attiecīgā informācija situācijā ir konceptuāli svarīga vai bez tās situācija var pastāvēt pat tad, ja attiecīgā informācija tieši neparādās teikumā, FN pieejā tiek šķirtas centrālās, perifērās un ārpustematiskās semantiskās lomas (Ruppenhofer et al. 2016). Informācija, ko sniedz ārpustematiskās semantiskās lomas, rāda attieksmes starp vairākām situācijām, bet vienas semantiskās grupas analīzē īpaši aktuālas ir centrālās un perifērās semantiskās lomas.

1. Centrālās semantiskās lomas raksturo situācijā konceptuāli nepieciešamus elementus, kas atšķir situāciju no citām (Ruppenhofer et al. 2016). Piemēram, tirdzniecības situācija nevar pastāvēt bez Pircēja, Pārdevēja, Preces.

2. Perifēro semantisko lomu nosauktā informācija neatšķir situāciju no cita tipa situācijām (tās var palīdzēt nošķirt vienu pārvietošanās situāciju no citas pārvietošanās situācijas), bet sniedz papildinformāciju par situāciju, tās apstākļiem. Šīs semantiskās lomas līdzīgi var raksturot jebkuru semantiski atbilstošu situāciju (Ruppenhofer et al. 2006). Gandrīz jebkuru situāciju var raksturot, piemēram, semantiskās lomas Laiks (šobrīd), Ātrums (*lēnām*), Cēlonis (*aiz garlaicības*) u. tml.:

Jānis šobrīd aiz garlaicības lēnām iet pa ielu.
Jānis šobrīd aiz garlaicības lēnām lasa grāmatu.

Izmantojot šo semantiskā apraksta pieeju, kopš 1997. gada tiek veidota mūsdienu angļu valodas leksiskā datubāze – Bērklijas *FrameNet* (turpmāk tekstā – BFN[1]), kurā iekļauti 1224 freimi, 13 685 leksiskās vienības un vairāk nekā 200 tūkstoši manuāli marķētu lietojuma piemēru[2].

Šobrīd šādi semantiski marķēti dati, kas nepieciešami gan valodniecības pētījumos, gan dabiskās valodas sapratnē un tekstradē, ir pieejami ne tikai angļu valodai, bet arī vairākām citām valodām, ieskaitot latviešu valodu (Gilardi, Baker 2018).

Latviešu valodas *FrameNet* dati

FrameNet-LV ir daļa no "Daudzslāņu valodas resursu kopas"[3] (FullStack-LV; Gruzitis et al. 2018a), kurā ievietotas no "Līdzsvarotā mūsdienu latviešu valodas korpusa" (LVK2018; Levāne-Petrova 2019) atlasītas rindkopas noteiktās reprezentatīvās proporcijās (60 % publicistikas, 20 % daiļliteratūras, 10 % normatīvie akti, 5 % Saeimas stenogrammas, 5 % citi teksti). Šo datu sintaktiskajā un semantiskajā marķēšanā izmantotas šādas dažādām valodām izmantojamas reprezentācijas: *Universal Dependencies* (UD; de Marneffe et al. 2021), FN, *PropBank* (Palmer et al. 2005) un *Abstract Meaning Representation* (AMR; Baranescu et al. 2013).

UD sintaktiskā reprezentācija latviešu valodas korpusam tiek automātiski atvasināta no sarežģītākās hibrīdās sintaktiskās marķēšanas sistēmas, kas ir izmantota "Latviešu valodas sintaktiski marķētajā korpusā" (Pretkalniņa et al. 2016). FN līmenis tiek marķēts uz UD pieejā sintaktiski marķētajiem datiem (Gruzitis et al. 2018b; Pretkalnina et al. 2018), savukārt no

1 *The Berkeley FrameNet lexical database*: http://framenet.icsi.berkeley.edu

2 https://framenet.icsi.berkeley.edu/fndrupal/current_status

3 Atvērtie dati pieejami: https://github.com/LUMII-AILab/FullStack

FN un sintaktiski marķētajiem datiem tiek atvasināts vispārīgākais *PropBank* semantiskā marķējuma slānis (Gruzitis et al. 2020). Papildus no zemākajiem marķējuma slāņiem daļēji automātiski var konstruēt AMR grafus.

Datu atlase

Latviešu valodas verbu lietojuma biežums ņemts vērā, jau veidojot FullStack-LV datu kopu – no 2000 LVK2018 biežāk lietotajiem latviešu valodas verbiem tika izveidots biežuma saraksts un, atlasot rindkopas, tika pārbaudīts, lai katrs no šiem verbiem FullStack-LV parādās vismaz 10 reizes.

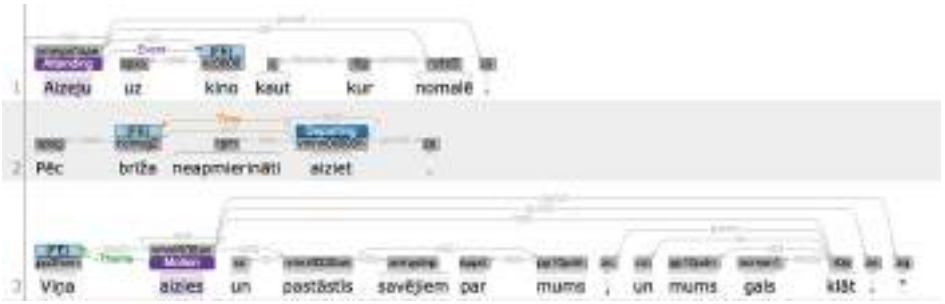
Tas pats verbu biežuma saraksts arī izmantots, tālāk no FullStack-LV izvēloties valodas materiālu semantiskajai marķēšanai. Priedēkļverbi, izņemot noliegtos verbus, tiek uzskatīti par atsevišķām vienībām, savukārt adjektīviski un substantīviski divdabji netiek šķirti no verbiem, kas ir to pamatā.

Deverbālu lietvārdu atlase bija sarežģītāka, jo lielāko daļu šo atvasinājumu nevar atlasīt no LVK2018 pēc morfoloģiskajām pazīmēm. Tāpēc deverbāli lietvārdi, kuru apkaimē parādās verbam raksturīgi paplašinātāji, pēc biežuma un sintaktiskās apkaimes tika konstatēti jau izveidotajā FullStack-LV. Tie atlasīti, gan ņemot vērā vārda sastāvu (regulāriem atvasinājumiem ar -šana, -šanās), gan "Latviešu valodas sintaktiski marķētajā korpusā" meklējot lietvārdus, kuriem piesaistīti, piemēram, sekundāri predikatīvi komponenti (*vēlme komponēt*) vai ar prepozicionālu konstrukciju izteikti apstākļi (*brauciens pie Buša*). Valodas materiāls, kas ar regulārajām izteiksmēm atrasts FullStack-LV, tiek pārskatīts, un deverbālie lietvārdi, kam ir verbam raksturīgi paplašinātāji un vismaz seši lietojuma piemēri, marķēti semantiski.

Datu marķēšana

FN līmenis tiek marķēts, izmantojot nevis sākotnējo rindkopu griezumu, bet gan konkrētā verba vai substantīva konkordances. Tas nozīmē, ka freimu marķēšanai no sintaktiski marķētajiem datiem tiek atlasīti teikumi, kuros parādās freimu marķēšanai izvēlētais verbs vai substantīvs (sk. verba *aiziet* leksiskās vienības 1. att.). Līdz ar to viens un tas pats teikums var parādīties vairākās šādās konkordanču kopās. Pēc tam teikumi un to marķējums atkal tiek sapludināti oriģinālajās rindkopās (2. att. sk. teikumu, kurā sapludināts divu freimu – DEATH 'nāve' un EXPERIENCER FOCUSED EMOTION 'izjutēja emocijas' – marķējums). Šāda pieeja nepieciešama, lai valodnieks, kas marķē freimus, vienkopus redz visus teikumus ar izvēlēto vārdu un konsekventāk var novērtēt, kurai nozīmei kurš freims vislabāk atbilst.

Tā kā perifērie un ārpustematiskie FE vairāk raksturo tekstā aprakstīto situāciju kopumā, nevis konkrētās leksiskās vienības leksisko nozīmi, latviešu valodas datos tiek marķēti galvenokārt centrālie FE. Perifēros FE marķē tikai gadījumos, ja leksisko vienību raksturo FE Laiks un Vieta, kā arī tad, ja perifērie FE kādai leksiskās vienības nozīmei ir specifiski, piemēram, verba *atgriezties* apkaimē teikumā *Jānis Vasiļonoks nesen atgriezās* mājās no tāla ceļojuma



1. attēls. Verbs *aiziet* konkordanču griezumā un tā semantiskais marķējums.

2. attēls. Teikums *Trešdienas vakarā 79 gadu vecumā mūžībā aizgājis tautā mīlētais dzejnieks Imants Ziedonis*, kurā apvienots divu freimu marķējums.

blakus centrālajiem FE Tēma (*Jānis Vasiļonoks*) un Mērķis (*mājās*) tiek marķēts arī perifērais FE Avots (*no ceļojuma*).

FN marķēšanai tiek izmantota datu marķēšanas platforma *WebAnno*[4] (Eckart de Castilho et al. 2016), kurai ir tīmekļa saskarne, plašas konfigurēšanas un pielāgošanas iespējas, kā arī ir piešķirts CLARIN pētnieciskās infrastruktūras atbalsts.

Semantiski tiek marķēts teikums, kas jau iepriekš ir sintaktiski marķēts UD pieejā (sk. 1. un 2. attēlu):

1. tiek atzīmēta marķējamā leksiskā vienība, tai izvēlas atbilstošo freimu (sk. freimus dažādām verba *aiziet* nozīmēm: ATTENDING ‘apmeklēšana’, DEPARTING ‘došanās prom’, MOTION ‘kustība’);
2. no konkrētā freima centrālo semantisko lomu saraksta izvēlas teikumā realizētās centrālās semantiskās lomas (sk. adverbiālu paplašinātāju *uz kino*, kam atbilst centrālā semantiskā loma *Event* ‘notikums’, 1. teikumā un teikuma priekšmetu *viņa*, kam atbilst centrālā semantiskā loma *Theme* ‘tēma’ 3. teikumā);
3. kur nepieciešams, pievieno perifērās semantiskās lomas (sk. adverbiālu paplašinātāju *pēc brīža*, kam atbilst perifērā semantiskā loma *Time* ‘laiks’).

4 https://webanno.github.io/webanno/

Marķējot semantiskā loma tiek piešķirta vienam konkrētam vārdam teikumā, bet, tā kā marķējums ir veikts kā nākamais līmenis pēc sintaktiskā marķējuma, visu frāzi, kurā ietilpst semantiskā loma, var iegūt automātiski, izvēloties visus no konkrētā vārda atkarīgos vārdus sintaktiskajā struktūrā: sk., piemēram, 2. attēlā – FE atzīmēts uz vārdformas *Imants*. Šai vārdformai sintaktiski ir piesaistītas citas vārdformas, un kopā veidojas frāze *tautā mīlētais dzejnieks Imants Ziedonis*. Pilnu frāzi plānots rādīt lietojuma piemēros FrameNet-LV.

No vienas puses, tas atvieglo marķēšanas procesu, jo valodniekam nav jāiezīmē konkrētās frāzes robežas, kas paildzinātu procesu. No otras puses, tas prasa noteiktas priekšzināšanas par to, kā veidojas sintaktiskā struktūra pielikuma, vienlīdzīgu teikuma locekļu un citos gadījumos, lai lomu piešķirtu tam vārdam, kam sintakses reprezentācijā būs pakārtoti visi pārējie frāzes vārdi (piem., atzīmējot FE, jāzina, ka frāzes *Imants Ziedonis* sintaktiskajā struktūrā otrā vārdforma tehniski ir pakārtota pirmajai, nevis otrādi, sk. 2. att.).

Freimu semantikā leksiskā vienība var sastāvēt arī no vairākiem vārdiem, tāpēc, veidojot latviešu valodas FN korpusu, atsevišķos gadījumos tiek izmantota iespēja norādīt, ka freima leksiskā vienība sastāv no vairākām vārdformām, jo verbs veido frazeoloģisku vienību, kurai atbilst kāds noteikts freims (sk. 2. att.: frazeoloģisms jeb leksiskā vienība *aiziet mūžībā*, kuras nozīmei atbilst freims DEATH ‘nāve’).

Sastatot latviešu valodas vārdu nozīmes ar BFN sistēmā piedāvātajiem freimiem, atsevišķos gadījumos rodas grūtības piemeklēt latviešu valodas verbam atbilstošu BFN freimu, jo:

1. BFN daudziem jēdzieniem vēl nav definēti freimi, piem., nozīmēm *veltīt, balsot, ziedēt*;
2. nav pietiekami precīzu atbilsmju starp angļu valodas un latviešu valodas jēdzieniem – latviešu valodas verba nozīmei atbilst angļu valodas vārdu savienojums (*kļūdīties*, *maldīties* ‘to be wrong’; ‘to make a mistake’) vai otrādi (*krist ģībonī*, *zaudēt samaņu* ‘to faint’), atšķiras latviešu un angļu verba nozīmes vispārinājuma pakāpe (*braukt* ‘to use vehicle’ – latviešu verba nozīmē nav ietverta informācija, vai persona vada transportlīdzekli vai ir tā pasažieris, bet angļu verbos šis nozīmes elements ir ietverts – attiecīgi verbi iekļaujas freimos RIDE VEHICLE ‘braukt (kā pasažierim) transportlīdzeklī’ vai OPERATE VEHICLE ‘vadīt transportlīdzekli’;
3. latviešu valodas verba nozīmei atbilst nomināls BFN freims, jo verba nozīme angliski izteikta ar palīgverbu un adjektīvu, piem., – *piedzerties* ‘to get drunk’ ietilpst freimā INTOXICATION ‘saindēšanās’.

Šādos gadījumos tiek meklēti papildu risinājumi, piem., nozīmes, kurām trūkst freimu, tiek marķētas kā UNDEFINED un tiek apkopotas, lai FN sistēmā var ieteikt jaunu freimu izveidi; ja verba nozīmei atbilst nomināls freims, tā tiek aprakstīta ar nominālu freimu (Nešpore-Bērzkalne et al. 2018).

Latviešu valodas FN korpusa tīmekļvietne

Līdz šim latviešu valodas korpusa semantiskajā marķēšanā ir izmantoti 570 BFN freimi, marķēts 2900 leksisko vienību (t. sk. 325 lietvārdi; vidēji 5,1 leksiskā vienība uz freimu) un gandrīz 26 000 lietojuma piemēru (vidēji 8,9 piemēri uz leksisko vienību)[5]. Lai šie dati pēc iespējas ērtāk būtu izmantojami lingvistiskiem pētījumiem, LU Matemātikas un informātikas institūtā (LU MII) ir izveidota FrameNet-LV tīmekļvietne[6], kurai par paraugu ņemta BFN vietne[7] (sk. 3. attēlu, kur norādīta freima ASSISTANCE 'palīdzība' definīcija, kā arī skaidroti centrālie freima elementi).

Freimus un leksiskās vienības var apskatīt kā sarakstu vai meklēt konkrētu freimu vai leksisko vienību. Ja izvēlas freimu, var aplūkot visas leksiskās vienības, kas marķētas, izmantojot šo freimu (sk. 5. att.).

Ja izvēlas leksiskās vienības, var redzēt visus freimus, ar kuriem katra leksiskā vienība marķēta (sk. 6. att.). Sadaļā "Leksiskās vienības" var atlasīt konkrētu vārdšķiru, piemēram, tikai lietvārdus (ierakstot meklētājā *noun*, tiek atlasīts lietvārdu saraksts).

Par konkrētu leksisko vienību pieejama šāda informācija:

1. visi tās lietojuma piemēri ar tajos marķētām semantiskajām lomām (sk. 4. attēla labo pusi);
2. pārskats par to, kādi FE ir šai nozīmei un kā tie realizējas gramatiski UD pieejā (sk. 7. attēlu);
3. pārskats par to, kādi valences modeļi jeb šabloni (visas semantisko lomu gramatiskās realizācijas kombinācijas, kas marķētas piemēros) veidojas šai nozīmei (sk. 8. attēlu).

FE gramatiskā realizācija norādīta saskaņā ar UD pieejā izmantoto vārdšķiru klasifikāciju[8] un definētajām universālajām sintaktiskajām attieksmēm[9], sk., piem., vārdšķiras PRON 'vietniekvārds', PROPN 'īpašvārds', VERB 'darbības vārds' un sintaktiskās attieksmes *nsubj* 'teikuma priekšmets', *obj* 'tiešais papildinātājs', *iobj* 'netiešais papildinātājs' 7. un 8. att. Atsevišķos gadījumos norādīti arī apakštipi, piem., *nsubj:pass* – teikuma priekšmets teikumos ar izteicēju ciešamajā kārtā.

5 Atvērtie dati pieejami: https://github.com/LUMII-AILab/FullStack/tree/master/FrameNet

6 http://framenet.korpuss.lv/

7 https://framenet.icsi.berkeley.edu/fndrupal/frameIndex

8 https://universaldependencies.org/u/pos/index.html

9 https://universaldependencies.org/u/dep/index.html

3. attēls. Freima ASSISTANCE 'palīdzība' definīcija un centrālo FE uzskaitījums un skaidrojums BFN tīmekļvietnē.

4. attēls. FrameNet-LV tīmekļvietne: kreisajā pusē – leksiskā vienība *palīdzēt* (freims ASSISTANCE 'palīdzība'), marķēto FE pārskats; labajā pusē – lietojuma piemēri, kuros ar melnu atzīmēts marķētais verbs, ar citām krāsām – FE ar tā paplašinātājiem, kas vizuāli ierāmēti.

5. attēls. Leksiskās vienības freimā ASSISTANCE 'palīdzība' pie katras iekavās norādīts marķēto piemēru skaits.

6. attēls. Leksiskās vienības *palikt* dažādos freimos.

Freima elementi	Piemēri	Realizācija
Helper	65	nsubj PRON (33) nsubj NOUN (22) nsubj PROPN (5) nsubj X (2) obj NOUN (1) nsubj VERB (1) nsubj:pass PRON (1)
Goal	56	ccomp VERB (53) dep VERB (1) iobj VERB (1) acl VERB (1)
Benefited_party	45	iobj NOUN (23) iobj PRON (17) iobj PROPN (2) iobj NUM (1) iobj VERB (1) ccomp VERB (1)
Focal_entity	4	obl NOUN (3) iobj NOUN (1)
Time	1	obl PRON (1)

7. attēls. Leksiskās vienības *palīdzēt* (ASSISTANCE ‘palīdzība’) FE, to biežums un gramatiskā realizācija.

Piemēri	Valences šabloni			
24	Goal	Helper		
11	ccomp VERB	nsubj PRON		
10	ccomp VERB	nsubj NOUN		
1	dep VERB	nsubj NOUN		
1	ccomp VERB	nsubj X		
1	ccomp VERB	nsubj:pass PRON		
20	Benefited_party	Goal	Helper	
7	iobj NOUN	ccomp VERB	nsubj PRON	
5	iobj PRON	ccomp VERB	nsubj PRON	
3	iobj NOUN	ccomp VERB	nsubj NOUN	
1	iobj PROPN	ccomp VERB	nsubj PRON	
1	iobj NOUN	iobj VERB	nsubj NOUN	
1	iobj NOUN	ccomp VERB	nsubj PROPN	
1	iobj NOUN	acl VERB	nsubj PRON	
1	iobj PRON	ccomp VERB	nsubj X	

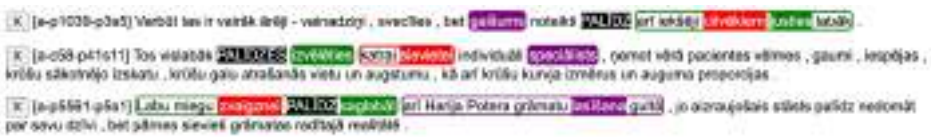
8. attēls. Leksiskās vienības *palīdzēt* (ASSISTANCE ‘palīdzība’) semantiskās un gramatiskās valences modeļu piemērs.

7. attēlā atspoguļots, kādi FE, cik reižu parādījušies marķētajos piemēros, kā arī norādīta realizēto FE leksēmu gramatika. Piem., FE *Helper* ‘palīgs’ realizēts kopumā 65 lietojuma piemēros, bet visbiežāk (33 reizes) tas izteikts ar vietniekvārdu teikuma priekšmeta funkcijā.

Lietojuma piemērus var atlasīt visus kopā vai kārtot, noņemot nost lieko, rādīt tikai teikumus, kuros ir konkrēti FE konkrētā gramatiskajā realizācijā, noteiktas FE kombinācijas u. tml. Piemēram, visi teikumi, kuros parādās FE *Helper* ‘palīgs’, kas gramatiski izteikts ar īpašvārdu teikuma priekšmeta funkcijā (sk. 9. att.). Turpinot piemēru atlasi, var nodzēst visus teikumus vai izvēlēties un pievienot vēl kādas kombinācijas.

Visas kombinācijas (3 piemēri), kur realizēti FE *Benefited party* ‘ieguvējs’ (sugasvārds netiešā papildinātāja funkcijā), *Goal* ‘mērķis’ (darbības vārds *ccomp* funkcijā (atbilstoši latviešu valodas sintakses teorijai – sekundāri predikatīvs komponents)), *Helper* ‘palīgs’ (sugasvārds teikuma priekšmeta funkcijā), redzamas 10. attēlā.

9. attēls. Ar īpašvārdu teikuma priekšmeta funkcijā izteikts FE *Helper* ‘palīgs’ leksiskās vienības *palīdzēt* apkaimē.



10. attēls. Viens gramatiskās valences modelis FE *Benefited_party* ‘ieguvējs’, *Goal* ‘mērķis’, *Helper* ‘palīgs’.

FN korpusa tīmekļvietne piedāvā datus aplūkot dažādos griezumos, piemēram, var salīdzināt semantisko lomu realizāciju pie konkrētiem verbiem un deverbāliem lietvārdiem – freimā PROTECTING ‘aizsardzība’ ietilpst gan verbs *aizsargāt*, gan arī lietvārds *aizsardzība*. Šī freima centrālās semantiskās lomas ir *Protection* ‘aizsargs’, *Asset* ‘manta’ un *Danger* ‘apdraudējums’. Verbam tās tipiski realizējas kā teikuma priekšmets, tiešais papildinātājs un netiešais papildinātājs, piemēram, fragmentā *12 bruņumašīnas var kādu zemes pleķi aizsargāt*. Savukārt deverbālā lietvārda *aizsardzība* apkaimē parasti realizējas tikai lomas *Asset* ‘manta’ (*nodarbināto*) un *Danger* ‘apdraudējums’ (*pret risku*): *nodarbināto aizsardzība pret darba vides trokšņa radīto risku*.

Tāpat var aplūkot, kādas semantiskās lomas parādās leksiskās vienības apkaimē, kā tās realizētas sintaktiski, kāda ir to tipiskā secība un ar ko skaidrojamas atkāpes no tipiskās secības. Plašais marķēto verbu un to lietojuma piemēru klāsts ļauj pētīt arī kādas noteiktas semantiskās grupas verbus.

Nobeigums

Semantiski marķēto datu kopa joprojām tiek papildināta, un FN korpusa tīmekļvietne tiek uzlabota, bet jau šobrīd skaidri redzams, ka šādi dati un piekļuve tiem ir nepieciešami gan valodu pētniekiem, gan valodas tehnoloģijām.

Marķēšana starptautiski pazīstamā un vienotā datu formātā ļauj latviešu valodas resursus integrēt starptautiski atzītās daudzvalodu datu kopās, vienotais datu formāts nodrošina, ka latviešu valodas dati kļūst pieejami un izmantojami arī valodu pētniekiem, kuri neprot latviešu valodu. Visus FN pieejā balstītos resursus un uz to pamata izstrādātās lietotnes kopējā daudzvalodu tīklā apvieno iniciatīva *Global FN*, kuras mērķis ir kopīgi FN balstīti pētījumi[10]. Viena no tās darbībām ir daudzvalodu FN pieejā marķēts paralēlais korpuss – viens teksts dažādās valodās (*TED Talk parallel corpus*; Torrent et al. 2018; Ohara 2020), kurā iekļauts arī marķēts teksts latviešu valodā.

Valodas pētniekiem šāds apjomīgs, ērti pieejams semantiski marķēts korpuss ļauj analizēt vārdu nozīmes un to lietojumu tekstā, bet dabiskās valodas apstrādē tā ir nozīmīga semantiskās marķēšanas treniņdatu kopa, ko var izmantot mašīntulkošanā, teksta sapratnē un tekstradē.

10 https://www.globalframenet.org/

Banarescu, Laura, Bonial, Claire, Cai, Su, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, Schneider, Nathan (2013). Abstract Meaning Representation for Sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178–186.

de Marneffe, Marie-Catherine, Manning, Christopher, Nivre, Joakim, Zeman, Daniel (2021). Universal Dependencies. *Computational Linguistics*, No. 47(2), pp. 255–308.

Fillmore, Charles J., Baker, Collin F. (2010). A Frames Approach to Semantic Analysis. Heine, B., Narrog, H. (eds.). *Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313–341.

Fillmore, Charles J., Johnson, Christopher R., Petruck, Miriam R. (2003). Background to FrameNet. *International Journal of Lexicography*, No. 16(3), pp. 235–250.

Fillmore, Charles, J. (1976). The need for a frame semantics in linguistics. Karlgren, H. (ed.). *Statistical Methods in Linguistics*. Stockholm: Scriptor, pp. 5–29.

Gilardi, Luca, Baker, Colin (2018). Learning to Align across Languages: Toward Multilingual FrameNet. *International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, pp. 13–22

Gruzitis, Normunds, Dargis, Roberts, Rituma, Laura, Nespore-Berzkalne, Gunta, Saulite, Baiba (2020). Deriving a PropBank Corpus from Parallel FrameNet and UD Corpora. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pp. 63–69.

Gruzitis, Normunds, Pretkalnina, Lauma, Saulite, Baiba, Rituma, Laura, Nespore-Berzkalne, Gunta, Znotins, Arturs, Paikens, Peteris (2018a). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pp. 4506–4513.

Gruzitis, Normunds, Nespore-Berzkalne, Gunta, Saulite, Baiba (2018b). Creation of Latvian FrameNet based on Universal Dependencies. *Proceedings of the International FrameNet Workshop (IFNW)*, pp. 23–27.

Levāne-Petrova, Kristīne (2019). Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Valoda: nozīme un forma,* 10, 131.–146. lpp.

Nešpore-Bērzkalne, Gunta, Saulīte, Baiba, Grūzitis, Normunds (2018). Latvian FrameNet: Cross-Lingual Issues. *Human Language Technologies – The Baltic Perspective,* IOS Press, Vol. 307, pp. 96–103.

Ohara, Kyoko (2020). Finding Corresponding Constructions in English and Japanese in a TED Talk Parallel Corpus using Frames-and-Constructions Analysis. *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pp. 8–12.

Palmer, Martha, Gildea, Daniel, Kingsbury, Paul (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, No. 31(1), pp. 71–106.

Rituma, Laura, Saulite, Baiba, Nešpore-Bērzkalne, Gunta (2019). Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis. *Valoda: nozīme un forma*, 10, 200.–216. lpp.

Ruppenhofer, Josef, Ellsworth, Michael, Petruck, Miriam R. L., Johnson, Christopher R., Baker, Collin F., Scheffczyk, Jan (2016). *FrameNet II: Extended Theory and Practice*. Available: https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf [accessed 05.11.2021.].

Torrent, Tiago Timponi, Ellsworth, Michael, Baker, Collin, Matos, Ely Edison da Silva (2018). The Multilingual FrameNet shared annotation task: A preliminary report. *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons (IFNW)*, pp. 62–68.

Summary

Latvian *FrameNet*

Baiba Saulīte, Gunta Nešpore-Bērzkalne, Laura Rituma, Viesturs Jūlijs Lasmanis, Normunds Grūzītis

Keywords: semantic annotation, frame semantics, semantic role, lexical unit, verb, deverbal derivative

This paper presents a FrameNet-annotated text corpus for Latvian language. Latvian FrameNet is a part of a FullStack corpus – medium-sized general-purpose multi-layered corpus, anchored in cross-lingual state-of-the-art syntactic and semantic representations: Universal Dependencies (UD), FrameNet and PropBank, as well as Abstract Meaning Representation (AMR). The FullStack has been designed considering the variety and balance of the corpus in terms of genres, domains, and lexical units.

For annotating the FrameNet layer in this corpus, we use the latest frame inventory of Berkeley FrameNet, while the annotation itself is done on top of the underlying UD layer. Thus, the annotation of frames and frame elements is guided by the dependency structure of a sentence, instead of the phrase structure. We strictly follow a corpus-driven approach, meaning that lexical units (verbs and deverbal derivatives) in Latvian FrameNet are created only based on the annotated corpus examples.

Currently, 570 Berkeley FrameNet frames have been used for semantic annotation of the Latvian FrameNet corpus, 2900 lexical units (average 5.1 lexical items per frame) and almost 26 000 usage examples (average 8.9 per lexical unit) have been tagged. To make this data available for linguistic research, the website of FrameNet-LV corpus has been created.