https://doi.org/10.22364/vnf.10.16

Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis

The grammar model of Latvian Treebank

Laura Rituma, Baiba Saulīte, Gunta Nešpore-Bērzkalne
Latvijas Universitātes Matemātikas un informātikas institūts
Mākslīgā intelekta laboratorija
Raiņa bulvāris 29, Rīga, LV-1050
E-pasts: *laura@ailab.lv, baiba@ailab.lv, gunta@ailab.lv*

Rakstā aplūkota „Latviešu valodas sintaktiski marķētā korpusa" izveide un attīstība, kā arī raksturots tajā izmantotais gramatikas modelis. Šis korpuss ir pirmais sintaktiski marķētais korpuss latviešu valodā, un šobrīd tā apjoms ir sasniedzis ap 13 000 sintaktiski marķētu teikumu. Lai iespējami precīzi aprakstītu latviešu valodas sintaktiskās konstrukcijas, korpusa marķēšanai ir izveidots hibrīds gramatikas modelis, kas ir balstīts atkarību sintaksē un papildināts ar frāzes struktūras sintakses elementiem. Ar frāzēm tiek attēlotas analītiskas vārdu formas un sintaktiskas vienības, starp kurām nav pakārtojuma sakars – salikti izteicēji, vienlīdzīgi teikuma locekļi, vairākvārdu nosaukumi u. c. Modeļa pamatā ir Lisjēna Tenjēra (*Lucien Tesnière*) ideja par sintaktiskajām vienībām, kas sastāv no vairākiem vārdiem, bet funkcionē kā viena sintaktiskā vienība.

Korpusa veidotāju izvēlētā marķēšanas pieeja un datu transformācijas iespējas nodrošina to, ka korpuss lietotājiem ir pieejams hibrīdā gramatikas modeļa formātā, kas ir ērtāks latviešu valodas sintakses izpētei, un universālo atkarību formātā, kas ir piemērotāks valodas tehnoloģijām.

Atslēgvārdi: latviešu valodas sintakse; sintaktiski marķēts korpuss; gramatikas modelis; atkarību sintakse; frāzes struktūras sintakse.

1. Ievads

Sintaktiski marķēts korpuss ir nepieciešams gan valodas tehnoloģijās – valodas automātiskās analīzes rīkos –, gan arī valodniecībā – dažādu sintaktisko konstrukciju pētīšanā, konkrētu leksēmu saistīšanās spējas analīzē u. c. Tagad arī latviešu valodai ir izveidots ap 13 000 teikumu liels sintaktiski marķēts korpuss – „Latviešu valodas sintaktiski marķētais korpuss" (turpmāk tekstā – LVTB)[1]. Tas ir automātiski morfoloģiski marķēts un manuāli sintaktiski marķēts tekstu kopums,

[1] LVTB ir izstrādāts ERAF praktiskas ievirzes pētījumu projektā „Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā" (1.1.1.1/16/A/219) sinerģijā ar ERAF pēcdoktorantūras pētniecības atbalsta projektu „No abstraktās nozīmes reprezentācijas līdz dabiskam teikumam un saistītam tekstam" (1.1.1.2/VIAA/1/16/188).

kurā katram teikumam ir kokveida struktūra un katram teikuma elementam – sintaktiskā loma. LVTB izveidē izmantots hibrīds gramatikas modelis – atkarību sintaksē balstīts gramatikas modelis, kas papildināts ar frāzes struktūras gramatikas elementiem.

LVTB tiek nodrošināts divos formātos – hibrīdā gramatikas modeļa formātā un universālo atkarību formātā. Hibrīdā gramatikas modeļa formātā korpuss pieejams[2] LINDAT/CLARIN Valodas pētniecības infrastruktūras centra (*Centre for Language Research Infrastructure in Czech Republic*) izveidotajā platformā dažādu valodu sintaktiski marķētu korpusu attēlošanai. Lai pievienotos starptautiskai iniciatīvai „Universālās atkarības"[3] (*Universal Dependencies*) un radītu iespēju LVTB izmantot arī daudzvalodu projektos, korpusa dati automātiski tiek transformēti arī universālo atkarību formātā, kas piedāvā universālas kategorijas un vadlīnijas, lai nodrošinātu saderīgu marķējumu līdzīgām konstrukcijām dažādās valodās (vairāk nekā 100 korpusu 70 dažādām valodām) (Nivre et al. 2016).

Lai izveidotu sintaktiski marķētu korpusu, vispirms bija jāizstrādā vai jāpielāgo gramatikas modelis, kā arī jāizvēlas marķēšanas rīks. Pēc tam, marķējot datus, tika attīstītas arī korpusa automātiskās pirmapstrādes iespējas, lai marķēšanas procesu varētu paātrināt. Šī raksta 2. nodaļā aprakstīts, kā LVTB veidojies un attīstījies laikā no 2007. gada līdz 2019. gadam un kādi teksti tajā tiek iekļauti. 3. nodaļā plašāk aprakstīts LVTB gramatikas modelis.

2. Latviešu valodas sintaktiski marķētā korpusa izveide un attīstība

„Latviešu valodas sintaktiski marķētā korpusa" izveide sākās 2007. gadā ar *Sem-Ti-Kamols* gramatikas modeļa izstrādi (Bārzdiņš, Grūzītis, Nešpore, Saulīte 2007). Par piemērotāko formālo modeli latviešu valodas sintakses teorijā aprakstīto parādību attēlošanai tika atzīts atkarībās balstīts hibrīds gramatikas modelis (sk. 3. nodaļu). Jau no paša korpusa izveides sākuma tika meklēti risinājumi, lai datus kaut daļēji varētu marķēt automātiski, tādēļ tika izveidots arī automātisks likumos balstīts gramatikas analizators, kas piedāvāja vairākus morfoloģiskās (Paikens 2007) un sintaktiskās analīzes variantus un spēja diezgan precīzi izanalizēt atsevišķas vārdkopas. Tomēr cilvēkam šajā rīkā bija jāizvēlas, kuru morfoloģiskās analīzes variantu katrai vārdformai apstiprināt, un analizators sintaktiski pareizi marķēja tikai nelielus teikuma fragmentus, pārējo teikuma struktūru vajadzēja marķēt manuāli, un tas padarīja marķēšanas procesu lēnu. Kā galvenais sintakses marķēšanas rīks izmantots Kārļa Universitātē izstrādātais sintakses koku redaktors *TrEd* (Hajič, Hladká, Pajas 2001), ar kuru saderīgais datu formāts PML tika pielāgots latviešu valodas vajadzībām (Pretkalniņa, Nešpore, Levāne-Petrova, Saulīte 2011a).

[2] Korpusa mājaslapā *sintakse.korpuss.lv* pieejama saite uz aktuālo laidienu (šobrīd 2.5. laidiens).

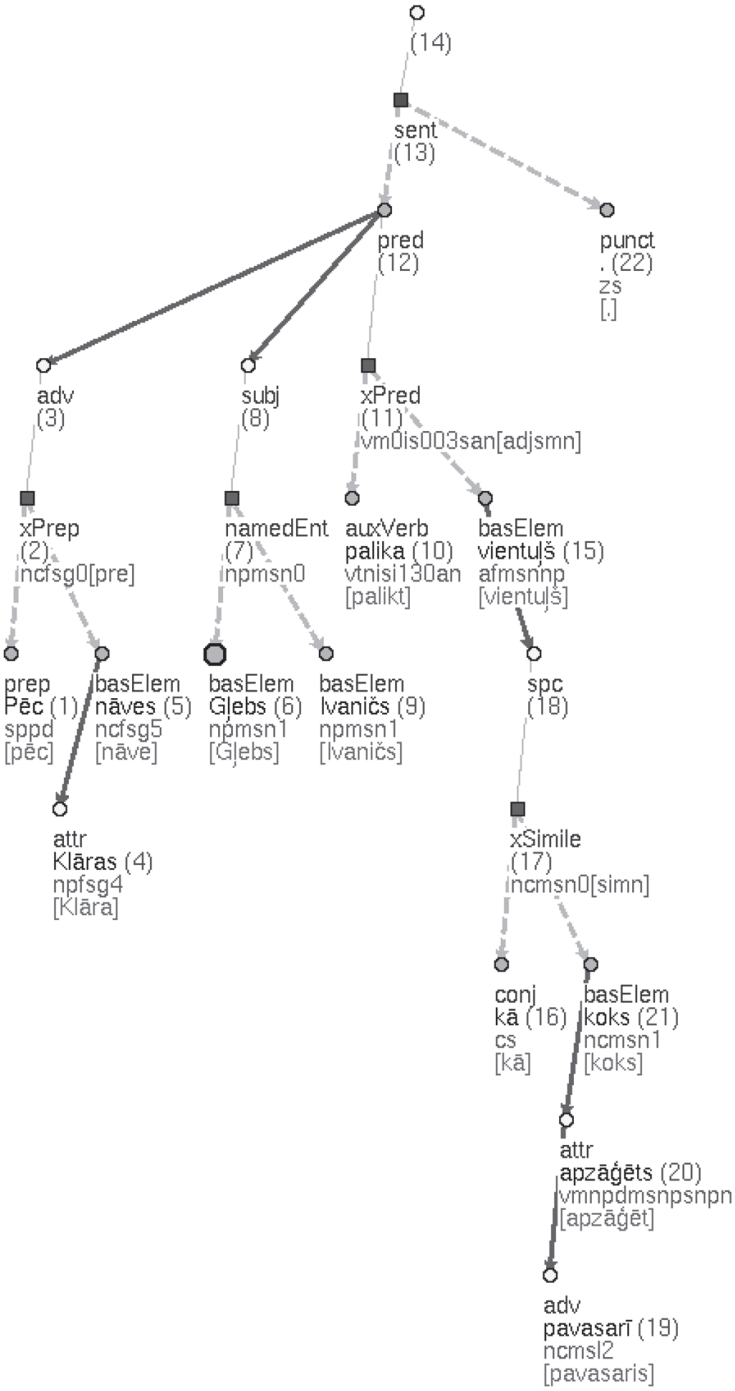
[3] Pieejams: http://universaldependencies.org/.

Šādā veidā vairāku gadu garumā tika marķēti ap 1500 teikumu (manuāli marķēja viens cilvēks), paralēli attīstot gramatikas modeli (Pretkalniņa, Nešpore, Levāne-Petrova, Saulīte 2011b; Pretkalniņa, Rituma 2012), un šāds apjoms jau deva iespēju izveidot parsētāju jeb automātisku sintaktisko analizatoru, kas tiktu apmācīts uz jau marķētajiem datiem un veiktu datu statistikā balstītus minējumus. Diemžēl šādi parsētāji datorlingvistikā ir pieejami vai nu atkarību, vai arī frāzes struktūras gramatikas modeļiem, bet ne tādam hibrīdajam gramatikas modelim, kāds tiek lietots LVTB. Tādēļ dati tiek transformēti atkarību formātā, izmantojot tos, tiek apmācīts parsētājs, un parsētāja iegūtais rezultāts ir atkarību formātā. Sākotnēji tas nedeva iespēju paātrināt datu manuālo marķēšanu hibrīdajā gramatikas modelī, bet tajā laikā attīstījās automātiskā morfoloģiskā analīze – tika izveidots automātiskais tagotājs jeb morfoloģiskais analizators (Paikens, Rituma, Pretkalniņa 2013), tādēļ datus varēja marķēt ātrāk un LVTB apjoms pieauga līdz apmēram 5000 teikumu 2014. gadā.

2016. gada beigās sākās jauns posms LVTB attīstībā, kad projektā „Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā" tika strauji attīstīts izmantotais gramatikas modelis, paātrināta marķēšana (marķē trīs cilvēki), izveidota likumos bāzēta datu transformācija uz universālo atkarību formātu (Pretkalniņa, Rituma, Saulīte 2016), apmācīts jauns parsētājs (Znotiņš 2016) un izveidota likumos bāzēta transformācijas sistēma atpakaļ uz hibrīdo gramatikas modeli (Grūzītis et al. 2018). Tas beidzot ir ļāvis iegūt automātiski marķētus datus hibrīdā gramatikas modeļa formātā. Šie dati joprojām ir tikai daļēji pareizi, bet tos manuāli labot un papildināt ir krietni vieglāk nekā iepriekš, un tas savukārt palīdz straujāk palielināt korpusa apjomu. 2019. gada nogalē LVTB apjoms ir sasniedzis ap 13 000 teikumu.

Laika gaitā ir mainījušies arī LVTB iekļaujamo tekstu atlases principi. Pirmais marķētais teksts bija 100 teikumi no Justeina Gordera grāmatas „Sofijas pasaule" (1996, no norvēģu valodas tulkojusi Brigita Šiliņa), jo tieši šos teikumus tolaik marķēja daudzās valodās, lai nodrošinātu sintaktiski marķētu paralēlo tekstu korpusu. Vēlāk tika marķēta arī latviešu oriģinālliteratūra – dažādu autoru stāsti. Laikā, kad korpusa dati bija nepieciešami sadarbības projektā ar ziņu aģentūru LETA, tika marķēta publicistika – ziņu raksti, arī intervijas.

2017. gadā tika formulēti daudzslāņu valodas resursu kopas izveides principi: visiem gramatiskā un semantiskā marķējuma līmeņiem tika izvēlēts viens un tas pats tekstu kopums – reprezentatīva apmēram 10 000 teikumu kopa, kas atlasīta no „Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa" (LVK2018; Levāne-Petrova 2019). Lai nodrošinātu datu kopas līdzsvarotību, šo korpusu veido no atsevišķām rindkopām (nevis pilniem tekstiem) tā, lai tajā būtu pārstāvēti 2000 LVK2018 biežāk lietotie verbi (Grūzītis et al. 2018). Īpaši izveidotā rindkopu šķirošanas un atlases rīkā tiek izvēlētas rindkopas no dažādu stilu tekstiem, ievērojot šādas proporcijas (tās tikai nedaudz atšķiras no LVK2018 proporcijām) – 60% periodika, 20% daiļliteratūra, 7% zinātniski teksti, 6% normatīvi teksti, 5% Saeimas stenogrammas un 2% citi teksti (Grūzītis et al. 2018).



1. attēls. Teikuma *Pēc Klāras nāves Gļebs Ivaničs palika vientuļš kā pavasarī apzāģēts koks* sintakses koks

3. „Latviešu valodas sintaktiski marķētajā korpusā" izmantotais gramatikas modelis

„Latviešu valodas sintaktiski marķētajā korpusā" ir izmantots hibrīds gramatikas modelis, kas ir balstīts atkarību sintaksē un papildināts ar frāzes struktūras gramatikas elementiem. Teikuma struktūra šajā pieejā tiek attēlota kā atkarību sintakses koks, kas papildināts ar dažādām frāžu veida konstrukcijām (sk. 1. attēlu). Atkarības attēlo atsevišķus vārdus vai vairākvārdu frāzes. Frāzes tiek izmantotas, lai attēlotu vairākvārdu formas un izteicienus, t. i., sintaktiskas vienības, ko veido analītiskas formas (piem., darbības vārda saliktā laika forma), vai sintaktiskas vienības, starp kurām nav pakārtojuma sakara (piem., vienlīdzīgi teikuma locekļi) (Pretkalniņa, Rituma, Saulīte 2018). Modelis ir balstīts uz Lisjēna Tenjēra (*Lucien Tesnière* 1959) ideju par sintaktiskām vienībām (franču val. *nucléus*), kas sastāv no vairākiem vārdiem, bet funkcionē kā viena sintaktiska vienība (sk. arī Nešpore, Saulīte, Bārzdiņš, Grūzītis 2010). Gan frāzei, gan tās sastāvdaļām (atsevišķiem vārdiem) var būt atkarīgie komponenti.

Šāds modelis vislabāk ļauj attēlot latviešu sintakses teorijā minētās analītiskās leksēmas (verba saliktie laiki, skaitļa vārdu savienojumi) un saliktos teikuma locekļus (vārdkopas analogi, vārdrindas analogi, vairākvārdu nosaukumi). Modelis ir piemērots arī plašākas sintaktiskās paradigmas izpratnes atainošanai, jo struktūras shēmā ļauj iekļaut izteicējus ar semantisko modificētāju, piem.:

(1) *Es gribu strādāt.*

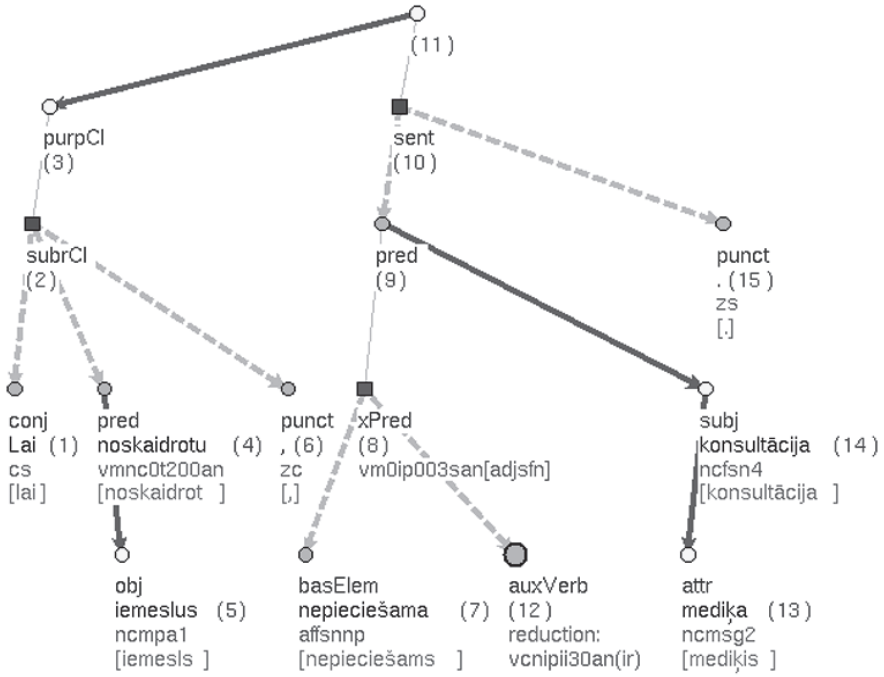
(2) *Man laimējās strādāt kopā ar viņu.*

Tāpat šādā modelī var norādīt šķīrumu starp vārda sakaru nosacītiem teikuma paplašinātājiem un teikuma sakaru nosacītiem paplašinātājiem, radot iespēju pēdējos attēlot kā elementus, kas atkarīgi no visa teikuma vai teikuma daļas. Modelis ļauj šķirt teikuma locekli, kas pakārtots visiem vienlīdzīgiem teikuma locekļiem, no teikuma locekļa, kas pakārtots tikai vienam no vienlīdzīgiem teikuma locekļiem (sk. 3. attēlu).

Hibrīdais gramatikas modelis paredz arī iespēju attēlot redukciju – ja teikumā ir izlaists kāds teikuma struktūras elements, kam pakārtoti citi teikuma locekļi, tiek izveidota īpaša redukcijas virsotne. Reducētu virsotņu attēlošana, ja virsotnei nav teikumā realizētu pakārtotu elementu, ir paredzēta tikai tad, ja tiek atjaunoti reducēti palīgverbi (sk. auxVerb 2. attēlā), reizēm reducēti semantiskie modificētāji:

(3) *Lai noskaidrotu iemeslus, [ir] nepieciešama mediķa konsultācija.*

Atjaunotajai virsotnei norāda trūkstošās vārdformas morfoloģijas tagu (sk. vārdformas *ir* morfoloģiskās pazīmes 2. attēlā), speciālā laukā var būt norādīta reducētā vārdforma, ja to var noteikt no konkrētā teikuma, bet, ja nevar, tad tikai tās morfoloģiskās pazīmes, kuras ir iespējams noteikt. Konkrētas vārdformas noteikšanā netiek izmantota informācija no blakus teikumiem.



2. attēls. Teikuma *Lai noskaidrotu iemeslus, [ir] nepieciešama mediķa konsultācija* sintakses koks ar saitiņas *ir* redukciju

3.1. Sintaktiskās lomas gramatikas modelī

Hibrīdajā gramatikas modelī sintaktiskās lomas (šajā rakstā šis termins lietots ar nozīmi 'apzīmējums') tiek izmantotas, lai apzīmētu katra teikuma elementa sintaktisko funkciju. Dažas no sintaktiskām lomām var būt attēlotas gan ar vienu vārdu, gan ar frāzi, bet citas sintaktiskās lomas var būt attēlotas tikai ar frāzi, kā, piem., palīgteikumi. Sintaktisko lomu saraksts redzams 1. tabulā.

Ar sintaktiskajām lomām tiek apzīmētas sintaktiskās funkcijas mūsdienu latviešu sintakses teorijā – izteicējs, teikuma priekšmets, papildinātājs, apzīmētājs un apstāklis, kā arī ārpusshēmas komponenti – situants, determinants un sekundāri predikatīvs komponents. Determinatīvi ārpusshēmas komponenti (determinants un situants) parasti ir saistīti ar visu predikatīvo vienību, nevis ar frāzi vai vienu vārdu. Determinants ir teikuma paplašinātājs datīvā, kas nosauc situācijas izjutēju vai īpašnieku un tiek attēlots kā visam teikumam pakārtots teikuma loceklis (Lokmane 2013, 758–759), piem.:

(4) ***Jums*** *patīk kārtība ne tikai apkārt, bet arī domās.*

Situants ir apstāklis, kas nosauc teikumā aprakstītās situācijas apstākļus un iesaistās teikumā neatkarīgi no vārda sakariem (Lokmane 2013, 760–761), piem.:

(5) ***Tagad*** *galvenais ir nomierināties pašiem, jo jūsu miers ir jūsu bērna veselības ķīla.*

Atkarību sintaktiskās lomas	
Pred	izteicējs
Subj	teikuma priekšmets
Attr	apzīmētājs
Obj	papildinātājs
Adv	apstāklis
Spc	sekundāri predikatīvs komponents
Sit	situants
Det	determinants
No	partikula

Frāžu sastāvdaļu sintaktiskās lomas	
Prep	prepozīcija
Mod	semantiskais modificētājs
auxVerb	palīgverbs
basElem	patstāvīgs vārds frāzē
Conj	saiklis
Punct	pieturzīme

Konstrukciju sintaktiskās lomas	
subjCl	teikuma priekšmeta palīgteikums
predCl	izteicēja palīgteikums
attrCl	apzīmētāja palīgteikums
objCl	papildinātāja palīgteikums
appCl	pielikuma palīgteikums
placeCl	vietas apstākļa palīgteikums
timeCl	laika apstākļa palīgteikums
manCl	veida apstākļa palīgteikums
degCl	mēra apstākļa palīgteikums
causCl	cēloņa apstākļa palīgteikums
purpCl	nolūka apstākļa palīgteikums
condCl	nosacījuma apstākļa palīgteikums
cnsecCl	seku palīgteikums
compCl	salīdzinājuma palīgteikums
cncesCl	pieļāvuma palīgteikums
motivCl	pamatojuma palīgteikums
quasiCl	relatīvais palīgteikums
ins	iespraudums, iestarpinājums
dirSp	tiešā runa

1. tabula. Sintaktiskās lomas

LVTB situants attēlots kā no visa teikuma atkarīgs teikuma loceklis. Tomēr minētās situanta nošķiršanas pazīmes pamatā ir semantiskas (verba valences nosacītas) un nav tik formāli aprakstāmas, lai varētu izveidot konsekventu marķējumu LVTB, tāpēc šobrīd šāda loma pamatā tiek izmantota gadījumos, kad situatīvs apstāklis attiecināts uz divām neatkarīgām teikuma daļām, piem.:

(6) ***Šoreiz**, salīdzinājumā ar citiem mačiem, komandai ir īsāks kandidātu saraksts un sastāvā iekļauti vien 18 futbolisti.*

Savukārt sekundāri predikatīva komponenta izpratne ir ļoti plaša (Lokmane 2013, 742–757; Nītiņa 2013, 810–819), un tā sastāvs var būt ļoti atšķirīgs – dažādas nominālas frāzes sekundāro predikātu nosaukšanai (7), salīdzinājuma konstrukciju salīdzinātājdaļa (8), divdabji (9):

(7) *Tētis sēdēja gultā **skumjš**, lai gan nekur nebija gājis, un darbojās ap ceļgaliem ...*

(8) *Šie Ziemeļāfrikāņu zēni ir pat sliktākā situācijā **nekā meitenes**.*

(9) *Beigās Antigone izdara pašnāvību* ***pakaroties****.*

Sintaktiskās lomas tiek lietotas, lai nosauktu teikumā iesaistītos palīgteikumus un citas sintaktiskās konstrukcijas, piem., iespraudumus, tiešo runu. Šajos gadījumos sintaktiskā loma tiek piešķirta virsotnei, kura ir izvērsta noteiktā pieturzīmju konstrukcijas frāzē (sk. raksta 3.2.3. nodaļu).

Arī partikulām LVTB ir sintaktiskā loma (**no)**. Ja partikula attiecas tikai uz vienu teikuma locekli, tā ir pakārtota konkrētai vārdformai (10) – partikula *tikai* iegūst sintaktisko lomu **no** un tiek pakārtota teikuma loceklim *Jānis.* Ja partikula nav saistīta ar kādu noteiktu teikuma locekli (11), tā tiek attēlota kā pieturzīmju konstrukcijas (teikuma vai teikuma daļas) sastāvdaļa.

(10) *Atnāca* ***tikai*** *Jānis.*

(11) ***Laikam*** *nav vērts jautāt sīkāk.*

Problēmas sintaktisko funkciju marķēšanā rada tas, ka līdz ar sekundāri predikatīva komponenta izpratnes paplašināšanu strukturālās sintakses ietvaros šajā grupā tiek iekļauts arvien plašāks parādību loks (Lokmane 2013, 742–757), tāpēc mainījusies arī tradicionālajā sintaksē (Ahero et al. 1962, 247–373) minēto atkarīgo teikuma locekļu (apzīmētāja, papildinātāja un apstākļa) izpratne, tiem atbilstošo parādību loks ir sašaurinājies un jāpārskata. Ilze Lokmane (2013, 761–765) vārda sakaru nosacītus paplašinātājus šķir pēc nozīmes, un tas rada atšķirības no teikuma locekļu izpratnes tradicionālajā gramatikā, tomēr trūkst plašāka teorētiska apraksta, kas apvienotu tradicionālās sintakses mantojumu ar mūsdienu sintakses izpratni. Piem., tradicionālajā gramatikā lietvārda ģenitīvs ar subjekta nozīmi pie lietvārda, kas darināts no verba, tiek uzskatīts par apzīmētāju (*suņa riešana*), bet jaunākās akadēmiskās gramatikas interpretācijā šādam subjekta ģenitīvam sintaktiskā funkcija nav norādīta, tikai teikts, ka tas ir paplašinātājs ar subjekta nozīmi.

Sekundāri predikatīvu komponentu aprakstā šādi paplašinātāji ar subjekta nozīmi nav iekļauti, tāpēc nav skaidrs, kādu sintaktisko lomu tiem piešķirt LVTB. Tāpat I. Lokmane (2013, 763) min iespēju paplašināt atribūtu izpratni un daļu tradicionālo apstākļu interpretēt kā atribūtus, tātad apzīmētājus (*ļoti priecīgs*, *jautri smieties*), tomēr paplašinātāju ar adverbiālu nozīmi jeb apstākļu aprakstā šī pieeja nav īstenota. LVTB pakārtotie teikuma locekļi – apzīmētājs, papildinātājs un apstāklis – marķēti pēc tradicionālās sintakses izpratnes, neiekļaujot šajās lomās parādības, kas iederas sekundāri predikatīvu komponentu grupā.

Ir sintaktiskās lomas, kas tiek piešķirtas tikai noteiktām frāžu sastāvdaļām:

- X-vārdos, ko veido no viena patstāvīga vārda un viena vai vairākiem palīgvārdiem, patstāvīgais vārds iegūst lomu **basElem**, bet palīgvārdam piešķir lomu atkarībā no frāzes tipa – **prep** (*uz skolu*), **mod** (*gribēja steigties*), **auxVerb** (*ir gājis*), **no** (*ne šovakar*). Ja x-vārds sastāv no vairākiem patstāvīgiem vārdiem, piem., pielikuma konstrukcija **xApp** (*māsa Anna*), tad visi frāzes elementi iegūst lomu **basElem**.
- Pieturzīmju konstrukcijas sastāv no viena **basElem** vai **pred**, vienas vai vairākām pieturzīmēm ar lomu **punct**, dažos gadījumos pieturzīmju konstrukcijas sastāvdaļa var būt saiklis **conj** vai arī partikula ar lomu **no**, piem., 3. attēlā redzamajā teikumā ir teikuma priekšmeta palīgteikums

subjCl, kuru veido pieturzīmju konstrukcija **subrCl** – tā sastāv no pieturzīmes **punct,** saikļa **conj** un predikāta **pred**, kuram pakārtots viss palīgteikuma koks.

- Sakārtojuma sastāvdaļas ir vienlīdzīgie teikuma locekļi vai teikuma daļas ar lomu **crdPart**, pieturzīmes ar lomu **punct** un/vai saikļi ar lomu **conj**.

3.2. Frāzes gramatikas modelī

Hibrīdajā gramatikas modelī tiek šķirtas triju veidu frāzes:

- **X-vārda** frāze sastāv no vairākiem vārdiem, bet visa frāze veic vienu sintaktisko funkciju. Ar x-vārdiem attēlo saliktus teikuma locekļus, piem., saliktu izteicēju, prievārdisku konstrukciju, nosaukumu u. c. vairākvārdu vienības.
- **Pieturzīmju konstrukcija** attēlo konstrukcijas, kas saistītas ar pieturzīmju lietojumu teikumā, piem., palīgteikumu, divdabja teicienu, izsauksmes vārdu, iespraudumu. PMC sastāv no vienas vai vairākām pieturzīmēm un tā bāzes elementa, kura dēļ tiek lietota pieturzīme.
- **Sakārtojuma konstrukcijas** attēlo sakārtojuma sakaru un sastāv no vienlīdzīgiem teikuma locekļiem vai vienlīdzīgām teikuma daļām, saikļiem un pieturzīmēm, kas atdala vienlīdzīgus teikuma locekļus vai vienlīdzīgas teikuma daļas.

X-vārdiem un sakārtojuma konstrukcijām tiek norādītas morfosintaktiskās pazīmes, kas raksturo, kā frāze iesaistās teikumā.

Katram frāžu veidam ir vairāki tipi, savukārt daļai x-vārdu ir arī apakštipi, kurus norāda frāzes morfoloģisko pazīmju aprakstā. Tas ļauj šķirt dažādas sintakses teorijā aprakstītās parādības.

3.2.1. X-vārdi

X-vārdi gramatikas modelī tiek izmantoti, lai attēlotu vairākvārdu frāzes, kas funkcionē kā viens vesels sintaktisks elements. Dažas frāzes sintakses teorijā ir precīzi un plaši aprakstītas – salikti izteicēji **xPred** (saliktas laika formas, sastata izteicēji, izteicēji ar semantisku modificētāju), prievārdiskas konstrukcijas **xPrep** (gan tradicionālas, gan konstrukcijas ar prievārdisku adverbu), skaitļa vārdu savienojumi **xNum**. Tomēr daļa x-vārdu gramatikās aprakstīti nepilnīgi, tādēļ, veidojot gramatikas modeli, to apraksts un dalījums apakštipos veidots, balstoties uz valodas materiālu (pilnu x-vārdu sarakstu sk. 2. tabulā).

- Salīdzinātājdaļa **xSimile** salīdzinājuma konstrukcijās gramatikās aprakstīta pamatā tikai pielīdzinājumos ar saikli *kā*, bet nav skatīts tās lietojums salīdzinājumos ar saikli *nekā.* Valodas materiāls rāda, ka var izveidot divus salīdzinātājdaļas apakštipus atkarībā no tā, vai frāze lietota konstrukcijā ar pielīdzinājuma nozīmi un struktūru (12) vai ar salīdzinājuma nozīmi un struktūru (13).

(12) *Bet tu jau esi tik vājš* ***kā niedre***.

(13) *Tas bija ļoti parasts stāsts, kurā es klausījos tikai ar vienu ausi un man likās, ka es to varētu izstāstīt labāk* ***nekā viņa***.

Turklāt frāzes tagā tiek īpaši marķētas gramatizējušās salīdzinājuma konstrukcijas *vairāk kā/nekā*, *mazāk kā/nekā* u. c., lai gan latviešu valodas sintakses teorijā pagaidām šādas konstrukcijas nav šķirtas.

- Vārdkopas analoga (Skujiņa 2007, 435) **subrAnal** izpratne sintakses teorijā nav precizēta, Intas Freimanes darbā „Valodas kultūra teorētiskā skatījumā" (Freimane 1993, 244–245) termins lietots attiecībā uz šauru valodas parādību loku (*mēs visi*, *tāds smieklīgs*, *kaut kas labs*), tomēr tur minētās pazīmes ļauj šī termina izpratni paplašināt un attiecināt to arī uz citiem vairākvārdu elementiem, ko mēdz saukt par sintaktiski vai jēdzieniski nedalāmiem vārdu savienojumiem (Ceplītis et al. 1989, 44–45), piem.:

(14) *viens no klausītājiem*

(15) *labākais no aktieriem*

- Ar **namedEnt** aprakstīti vairākvārdu nosaukumi, piem., cilvēka vārds *Jānis Bērziņš.* Kā frāze tiek attēloti tikai tie nosaukumi, kuru iekšējā struktūrā nav pakārtojuma sakara, piem., *Bērziņš Investment*. Ja nosaukumā ir pakārtojuma sakars, frāze netiek veidota, piem., *Latvijas Universitāte* netiek attēlots kā x-vārds.

X-vārds	Apraksts	Apakštips	Apraksts
xPrep	prievārdiska konstrukcija	[pre]	konstrukcija ar prepozitīvu prievārdu: *ar roku*
		[post]	konstrukcija ar postpozitīvu prievārdu: *naudas dēļ*
		[rel]	konstrukcija ar prievārdisku adverbu: *tuvu ābelei*
xPred	salikts izteicējs	[act], [pass]	saliktie laiki: *ir gājis*
		[subst], [adj], [pronom], [inf], [num], [adv]	sastata izteicēji – nomināli, pronomināli, adverbiāli un sastata izteicējs ar infinitīvu: *bija spēcīgs*
		[modal], [phase], [expr]	salikts izteicējs ar modificētāju: *gribēja skriet*
xNum	skaitļa vārdu savienojums: *divdesmit četri*		
xApp	pielikums	[agr]	saskaņots pielikums: *māsa Anna*
		[non]	nesaskaņots pielikums: *laikraksts „Diena"*
xSimile	salīdzinātājdaļa	[sim]	pielīdzinājums: *kā Jānis*
		[comp]	salīdzinājums: *nekā vakar*
xParticle	konstrukcijas ar partikulu (partikula attiecas uz noteiktu teikuma locekli un ir fiksētā pozīcijā)	[aff]	konstrukcija bez nolieguma: *ik rītu, kaut viens*
		[neg]	konstrukcija ar noliegumu: *ne vienmēr*

X-vārds	Apraksts	Apakštips	Apraksts
namedEnt	personvārdi, organizāciju nosaukumi: *Jānis Bērziņš*		
subrAnal	vārdkopas analogs	[vv]	vietniekvārdu savienojums: *mēs visi*, *tas pats*
		[ipv]	vietniekvārda un adjektīva vai adjektīviska divdabja savienojums: *tāds interesants*
		[skv]	vietniekvārda un skaitļa vārda savienojums: *abi divi*, *visi trīs*
		[set]	kopuma konstrukcija: *viens no viņiem*, *labākais no aktieriem*
		[sal]	gramatizējušās salīdzinājuma konstrukcijas: *tāds kā noraizējies*, *tāds kā rēgs*
		[part]	vairākvārdu partikula: *it kā*, *diez vai*
coordAnal	vārdrindas analogs: *trīs četri*		
unstruct	teksta primitīvu virknes bez gramatiskās struktūras: *per aspera ad astra, f(x) = sin x + |x| + 17*		
phraseElem	frazeoloģiska vienība, kas iekļaujas teikumā kā viens veselums: *[pabeidza] viens un divi, ka tavu almu māteri*		

2. tabula. X-vārdu tipi un apakštipi

- Pielikuma konstrukcija **xApp**, kurā iekļauti gan saskaņoti (16, 17), gan nesaskaņoti (18) pielikumi. Par šīs konstrukcijas iekšējiem sintaktiskajiem sakariem pastāv dažādi uzskati (Lokmane 2013, 766), bet frāze likās atbilstošākais veids, kā to marķēt LVTB.

(16) *māte daba*

(17) *māsa Anna*

(18) *laikraksts „Diena”*

Pamatā ar x-vārdiem tiek attēloti sintakses teorijā aprakstītie saliktie teikuma locekļi, bet x-vārdi ļauj modelī iekļaut arī tādas sintaktiskās parādības, kas vēl nav sīkāk aprakstītas, tomēr, pēc LVTB veidotāju domām, veido noteiktu vairākvārdu elementu tipu. Arī tad, ja kāda x-vārda interpretācija raisa diskusiju, šāds marķējums ļauj atrast līdzīgos gadījumus un veikt pētījumu, lai precizētu attiecīgās parādības izpratni.

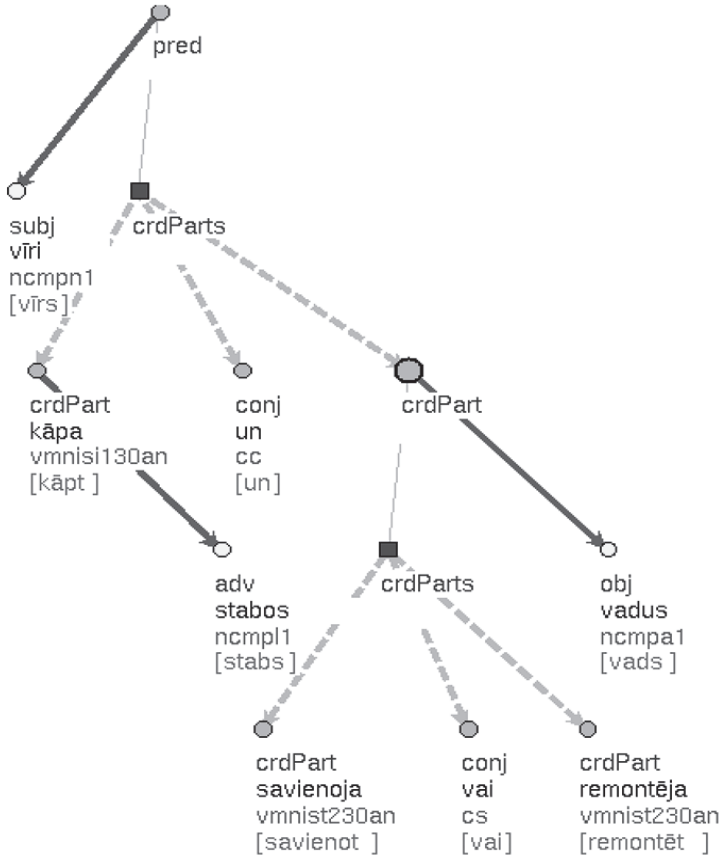
Modelis ļauj izmantot x-vārdu pieeju arī tādām parādībām, kas tradicionāli netiek uzskatītas par saliktiem teikuma locekļiem, bet kas gramatikas aprakstos tiek sauktas par vairākvārdu elementiem – vairākvārdu partikulām un saikļiem. Šis varētu būt viens no nākamajiem soļiem modeļa attīstīšanā – vairākvārdu partikulas šobrīd tiek marķētas kā viens no vārdkopas analoga **subrAnal** apakštipiem, tomēr tās neatbilst vārdkopas analoga izpratnei. Vairākvārdu elementiem ar palīgvārda funkciju būtu jāizveido savs x-vārda tips ar diviem apakštipiem – vairākvārdu

partikulām un vairākvārdu saikļiem. Vēl jāpēta un jāprecizē arī tādu vārdu savienojumu kā (19) iederība vārdkopas analoga kategorijā.

(19) *tāds kā noraizējies*

3.2.2. Sakārtojuma konstrukcijas

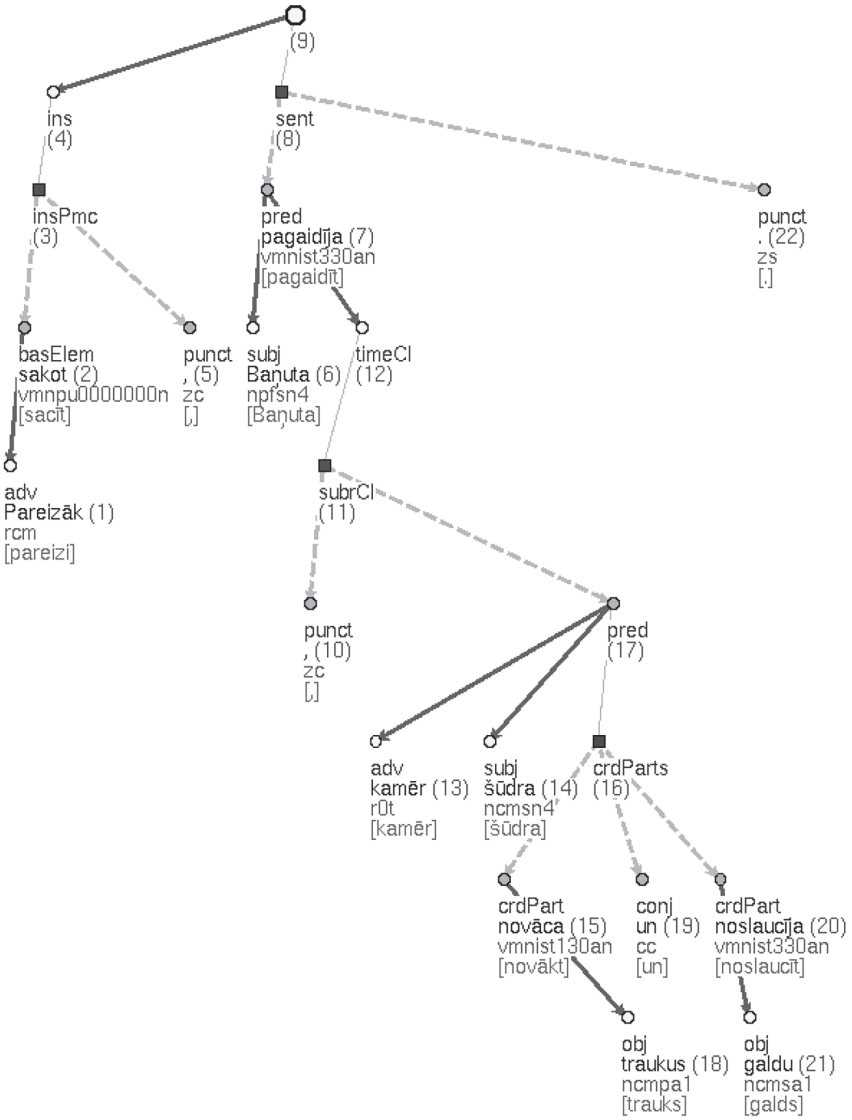
Sakārtojuma konstrukcijas tiek izmantotas, lai attēlotu sakārtojuma sakaru – vienlīdzīgus teikuma locekļus (**crdParts**) vai vienlīdzīgas teikuma daļas (**crdClauses**). Frāze ietver vienlīdzīgos teikuma locekļus, kā arī saikļus un pieturzīmes, kas tos atdala. Var attēlot arī divpakāpju sakārtojumu – gadījumus, kad kāds no vienlīdzīgiem teikuma locekļiem ir izvērsts plašāk divos vai vairākos vienlīdzīgos teikuma locekļos, piem., teikuma daļā .. *vīri kāpa stabos un savienoja vai remontēja vadus* izteicēji *savienoja* un *remontēja* abi kopā ir vienlīdzīgi ar izteicēju *kāpa* (sk. 3. attēlu).



3. attēls. Divpakāpju sakārtojuma konstrukcija. Sintakses koks teikuma daļai .. *vīri kāpa stabos un savienoja vai remontēja vadus*

3.2.3. Pieturzīmju konstrukcijas

Pieturzīmju konstrukcijas gramatikas modelī ir frāzes, kas ļauj piesaistīt pieturzīmes tam teikuma elementam vai teikuma daļai, kuras dēļ teikumā lietotas pieturzīmes. Līdz ar to arī teikuma beigu pieturzīme ar visu teikumu kopā veido pieturzīmju konstrukciju **sent**, divdabja teiciens kopā ar to atdalošajām pieturzīmēm veido **spcPmc**, tiešā runa kopā ar kolu un pēdiņām vai domuzīmēm veido **dirSpPmc** utt. 4. attēlā redzamas trīs pieturzīmju konstrukcijas – teikums **(sent)**, palīgteikums **(subrCl)** un iespraudums **(insPmc)**.



4. attēls. Pieturzīmju konstrukcijas teikumā. Teikuma *Pareizāk sakot, Baņuta pagaidīja, kamēr šūdra novāca traukus un noslaucīja galdu* sintakses koks

Šāds iedalījums pamatā atbilst sintakses teorijā (Lokmane 2013) aprakstītajām konstrukcijām, tomēr arī šeit modelis būtu vēl jāattīsta, lai panāktu konsekventu sekundāri predikatīvu konstrukciju marķēšanu, vienlaikus nepazaudējot teorijā aprakstītos konstrukciju tipus – daļa no konstrukcijām, kas ir sekundāri predikatīvas, nav šķirtas kā atsevišķi tipi, piem., savrupinājumi un paskaidrojošās vārdu grupas tiek veidotas zem atkarību lomas **spc** ar PMC frāzi **spcPmc**, tāpat kā divdabja teicieni, līdz ar to tās atrast korpusā nav tik vienkārši. Savukārt iespraudumu konstrukcija **insPmc** aptver gan nepredikatīvas, gan sekundāri predikatīvas, gan predikatīvas vienības, no kurām atšķirīgi marķēti tikai predikatīvi iespraudumi, norādot, ka to sastāvā ir predikatīva vienība **pred**, nevis patstāvīgs vārds **basElem**, kā pirmajos divos gadījumos. Pieturzīmju konstrukciju saraksts redzams 3. tabulā.

PMC	Apraksts
sent	teikums (koka sakne)
utter	izteikums (koka sakne)
mainCl	virsteikums, neatkarīga teikuma daļa
subrCl	palīgteikums
interj	izsauksmes vārds pieturzīmēs
spcPmc	divdabja teicieni, savrupinājumi, paskaidrojošas vārdu grupas
insPmc	iespraudumi un iestarpinājumi pieturzīmēs
particle	partikula pieturzīmēs
dirSpPmc	tiešās runas pieturzīmju konstrukcija
address	uzruna
quot	pēdiņu lietojums, kas nav saistīts ar tiešo runu – citāts, nosaukums

3. tabula. Pieturzīmju konstrukciju tipi

4. Secinājumi

„Latviešu valodas sintaktiski marķētā korpusa" izveidē izmantotais hibrīdais gramatikas modelis palīdz attēlot latviešu sintakses teorijā aprakstītās vārdkopas, analītiskās leksēmas un saliktos teikuma locekļus. Izveidotie frāžu tipi un apakštipi aptver lielu daļu sintakses teorijā aprakstīto parādību. Atšķirībā no atkarību gramatikas modeļiem tas ļauj piešķirt vienu sintaktisko funkciju vairākvārdu frāzei, kā arī ļauj šķirt visas frāzes atkarīgos no šīs frāzes atsevišķa elementa atkarīgajiem. Sīki izstrādātais frāžu sastāvdaļu marķējums ļauj transformēt datus uz dažādiem atkarību modeļiem, nemainot oriģinālā modeļa marķēšanas principus un datus.

LVTB ir ļoti vērtīgs pētījumu avots, korpusa apjoms ir pietiekams, lai to varētu izmantot dažādiem latviešu valodas sintakses pētījumiem. Īpaši noderīgi būtu pētījumi par nepilnīgi aprakstītām, bet plaši sastopamām sintaktiskajām konstrukcijām latviešu valodā, piem., vārdkopas analogiem, pielikuma konstrukcijām, reducētiem semantiskiem modificētājiem saliktos izteicējos. Korpusa izmantošana sintakses pētījumos ļautu gan novērtēt, gan uzlabot LVTB gramatikas modeli.

Turpmāk paredzēts palielināt LVTB apjomu, samazināt manuāli pieļautu kļūdu un nekonsekventa marķējuma gadījumu skaitu, kā arī attīstīt gramatikas modeli, izveidojot sekundāri predikatīvu komponentu apakštipus, ieviešot jaunu x-vārdu vairākvārdu saikļiem u. tml., lai pēc iespējas precīzāk aptvertu latviešu valodas gramatikā aprakstītās parādības.

Saīsinājumi

LVK2018	*Līdzsvarotais latviešu valodas tekstu korpuss*
LVTB	*Latviešu valodas sintaktiski marķētais korpuss*
PMC	pieturzīmju konstrukcija (*Punctuation Mark Construction*)
PML	datu formāts *Prague Markup Language*

Avots

Latviešu valodas sintaktiski marķētais korpuss. Pieejams: https://sintakse.korpuss.lv.

Literatūra

1. Ahero, Antonija et al. 1962. *Mūsdienu latviešu literārās valodas gramatika.* II. *Sintakse*. Rīga: Latvijas PSR Zinātņu akadēmijas izdevniecība.
2. Bārzdiņš, Guntis, Gruzītis, Normunds, Nešpore, Gunta, Saulīte, Baiba. 2007. Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*. Nivre, Joakim, Kaalep Heiki-Jaan, Muischnek Kadri, Koit Mare (eds.). Tartu: University of Tartu, 13–20.
3. Ceplītis, Laimdots, Rozenbergs, Jānis, Valdmanis, Jānis. 1989. *Latviešu valodas sintakse.* Rīga: Zvaigzne.
4. Freimane, Inta. 1993. *Valodas kultūra teorētiskā skatījumā*. Rīga: Zvaigzne.
5. Grūzītis, Normunds, Pretkalniņa, Lauma, Saulīte, Baiba, Rituma, Laura, Nešpore-Bērzkalne, Gunta, Znotiņš, Artūrs, Paikens, Pēteris. 2018. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. Calzolari, Nicoletta et al. (eds.). Miyazaki: European Language Recources Association (ELRA), 4506–4513.
6. Hajič, Jan, Hladká, Barbora, Pajas, Petr. 2001. The Prague Dependency Treebank: Annotation Structure and Support. *Proceedings of the IRCS Workshop on Linguistic Databases*. Bird, Steven, Liberman, Mark, Buneman, Peter (eds.). Philadelphia: University of Pensylvania, 105–114.
7. Levāne-Petrova, Kristīne. 2019. LVK2018: Līdzsvarotais latviešu valodas tekstu korpuss, tā nozīme gramatikas pētījumos. *Valoda: nozīme un forma.* 10. *Latvijas gramatiskā doma gadsimta gaitā.* Kalnača, Andra, Lokmane, Ilze (red.). Rīga: LU Akadēmiskais apgāds.
8. Lokmane, Ilze. 2013. Vienkārša teikuma formālā (strukturālā) organizācija. *Latviešu valodas gramatika.* Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: LU Latviešu valodas institūts, 710–766.
9. Nešpore, Gunta, Saulīte, Baiba, Bārzdiņš, Guntis, Grūzītis, Normunds. 2010. Comparison of the SemTi-Kamols and Tesnière's Dependency Grammars.

Proceedings of the 4th International Conference on Human Language Technologies – the Baltic Perspective. *Frontiers in Artificial Intelligence and Applications*. 219. Skadiņa, Inguna, Vasiļjevs, Andrejs (eds.). Amsterdam: IOS Press, 233–240.

10. Nītiņa, Daina. 2013. Vienkārša teikuma vai salikta teikuma komponentu paplašinājumi (paplašinātājas struktūras). *Latviešu valodas gramatika.* Nītiņa, Daina, Grigorjevs, Juris (red.). Rīga: LU Latviešu valodas institūts, 801–829.
11. Nivre, Joakim, Marneffe, Marie-Catherine de, Ginter, Filip, Goldberg, Yoav, Hajič, Jan, Manning, Christopher D., McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo, Silveira, Natalia, Tsarfaty, Reut, Zeman, Daniel. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC).* Calzolari, Nicoletta, et al. (eds.). Paris: European Language Recources Association (ELRA), 1659–1666.
12. Paikens, Pēteris. 2008. Lexicon-based morphological analysis of Latvian language. *Proceedings of the 3rd Baltic Conference on Human Language Technologies – the Baltic Perspective. Frontiers in Artificial Intelligence and Applications.* Čermák, František, Marcinkevičienė, Rūta, Rimkutė, Erika, Zabarskaitė, Jolanta (eds.). Vilnius: Vytautas Magnus University, Institute of the Lithuanian Language, 235–240.
13. Paikens, Pēteris, Rituma, Laura, Pretkalniņa, Lauma. 2013. Morphological analysis with limited resources: Latvian example. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*. Oephen, Stephan, Hagen, Kristin, Johannesen, Janne Bondi (eds.). Oslo: Linköping University Electronic Press, 267–277.
14. Pretkalniņa, Lauma, Nešpore, Gunta, Levāne-Petrova, Kristīne, Saulīte, Baiba. 2011a. A Prague Markup Language Profile for the SemTi-Kamols Grammar Model. *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA)*. Pedersen, Bolette Sandford, Nešpore, Gunta, Skadiņa, Inguna (eds.). Riga: Northern European Association for Language Technology (NEALT), 303–306.
15. Pretkalniņa, Lauma, Nešpore, Gunta, Levāne-Petrova, Kristīne, Saulīte, Baiba. 2011b. Towards a Latvian Treebank. *Actas del 3 Congreso Internacional de Lingüística de Corpus. Tecnologias de la Información y las Comunicaciones: Presente y Futuro en el Análisis de Corpus*. Candel Mora, Miguel Angel, Carrió Pastor, Maria Luisa (eds.). Valnecia: Editorial Universitat Politecnica de Valencia, 119–127.
16. Pretkalniņa, Lauma, Rituma, Laura. 2012. Syntactic Issues Identified Developing the Latvian Treebank. *Proceedings of the 5th International Conference on Human Language Technologies – the Baltic Perspective. Frontiers in Artificial Intelligence and Applications*. 247. Tavast, Arvi, Muischnek, Kadri, Koit, Mare (eds.). Amsterdam: IOS Press, 185–192.
17. Pretkalniņa, Lauma, Rituma, Laura, Saulīte, Baiba. 2016. Universal Dependency Treebank for Latvian: A Pilot. *Proceedings of the 7th International Conference on Human Language Technologies – the Baltic Perspective. Frontiers in Artificial Intelligence and Applications*. 289. Skadiņa, Inguna, Rozis, Roberts (eds.). Amsterdam: IOS Press, 136–143.
18. Pretkalniņa, Lauma, Rituma, Laura, Saulīte, Baiba. 2018 Deriving enhanced Universal Dependencies from a hybrid dependency-constituency treebank. *Proceedings of the 21st International Conference "Text, Speech, and Dialogue"*

(TSD). Lecture Notes in Computer Science. 11107. Sojka, Petr, Horák, Aleš, Kopeček, Ivan, Pala, Karel (eds.). Brno: Springer-Verlag, 95–105.

19. Skujiņa, Valentīna (red.). 2007. *Valodniecības pamatterminu skaidrojošā vārdnīca.* Rīga: LU Latviešu valodas institūts.
20. Ten'er, Ljus'en. 1988. *Osnovy strukturnogo sintaksisa*. Gak, Vladimir (red.). Moskva: Progress. [Pirmizd.: Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.]
21. Znotiņš, Artūrs. 2016. Word embeddings for Latvian natural language processing tools. *Proceedings of the 7th International Conference on Human Language Technologies – the Baltic Perspective. Frontiers in Artificial Intelligence and Applications*. 289. Skadiņa, Inguna, Rozis, Roberts (eds.). Amsterdam: IOS Press, 167–173.

Summary

This paper describes the development of *Latvian Treebank* and its grammar model. This corpus is the first syntactically annotated corpus for Latvian, and currently contains approximately 13000 annotated sentences. A hybrid dependency-constituency model was developed in order to describe Latvian syntactic constructions as accurately as possible, augmenting the commonly used dependency grammars with phrase constructions for certain syntactic elements – analytical word forms and relations other than subordination. The grammar model is based on idea of a syntactic nucleus which is a functional syntactic unit consisting of content-words or syntactically inseparable units that are treated.

There are three kinds of phrase constructions in the Latvian Treebank grammar model: x-words, coordination and punctuation mark constructions. X-words are used for analytical forms, compound predicates, prepositional phrases etc. Coordination constructions are used for coordinated parts of sentences and coordinated clauses. Punctuation mark constructions are used to annotate different types of constructions that require the punctuation in the sentence.

The chosen annotation approach and data transformation systems ensure that the corpus is accessible to end users both in the hybrid dependency-constituency model suitable for research of syntactic phenomena in Latvian linguistic tradition, and in the Universal Dependencies multilingual model that is better suited for certain computational linguistics systems.

This work has received financial support from European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219 (Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian) in synergy with the grant agreement No. 1.1.1.2/VIAA/1/16/188 (From Abstract Meaning Representation to Natural Language Sentence and Coherent Text Generation).

Keywords: Latvian syntax; treebank; grammar model; dependency syntax; phrase structure grammar.