PostDoc Latvia

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020

EIROPAS SAVIENĪBA
Eiropas Reģionālās attīsītības fonds

IEGULDĪJUMS TAVĀ NĀKOTNĒ

CLARIN
Common Language Resources and Technology Infrastructure

Sintaktiski marķēta latviešu valodas korpusa gramatikas modelis

Gunta Nešpore-Bērzkalne, Laura Rituma, Baiba Saulīte

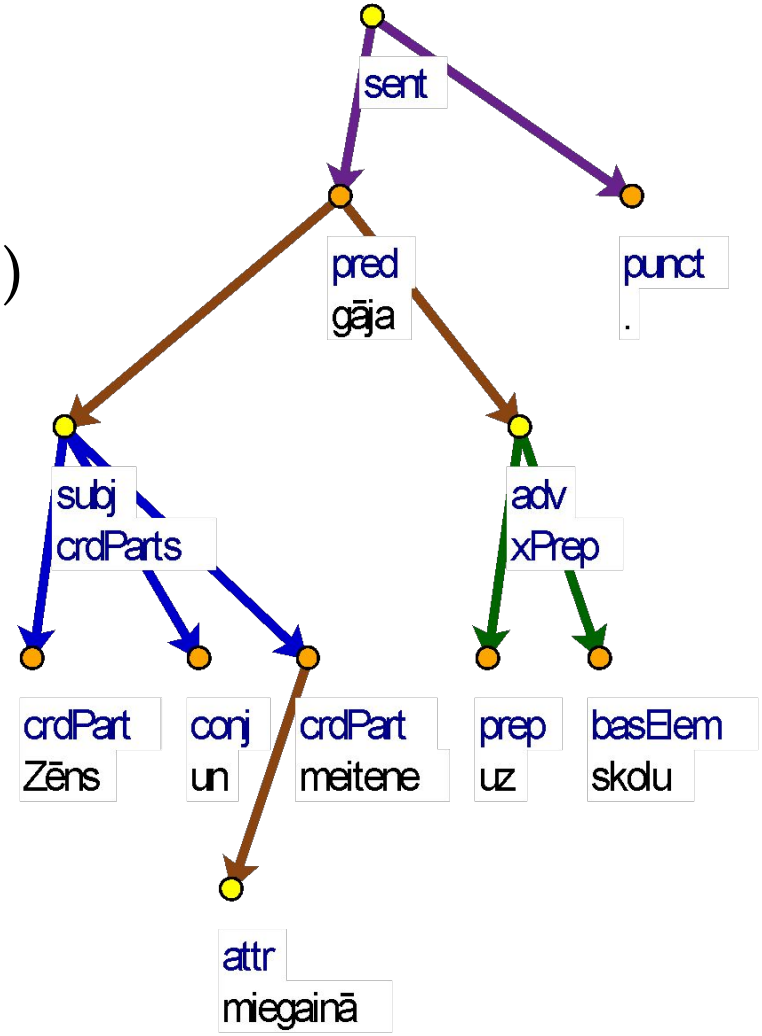
LU Matemātikas un informātikas institūts

Praktiskas ievirzes pētījumu proj. "Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā" (Nr.1.1.1.1/16/A/219)
CLARIN darbību Latvijā atbalsta ERAF projekts "Latvijas Universitāte un institūti Eiropas pētniecības telpā - ekselence, aktivitāte, mobilitāte, kapacitāte" (Nr. 1.1.1.5/18/I/016)

Mākslīgā intelekta laboratorija
LU MII

Kas ir korpuss?

- Sintaktiski (un morfoloģiski) marķētu tekstu kopums
- Katram teikumam:
  - kokveida struktūra
  - sintaktiskās lomas



Zēns un miegainā meitene gāja uz skolu.

Mākslīgā intelekta laboratorija
LU MII

Korpusa izmantojums

Valodu tehnoloģijās – valodas automātiskas analīzes rīkos

Valodniecībā – sintaktisko konstrukciju pētīšanā, piemēram:

- dažādu sintaktisko konstrukciju meklēšana

  Kādi verbi var veikt saitiņas funkciju?

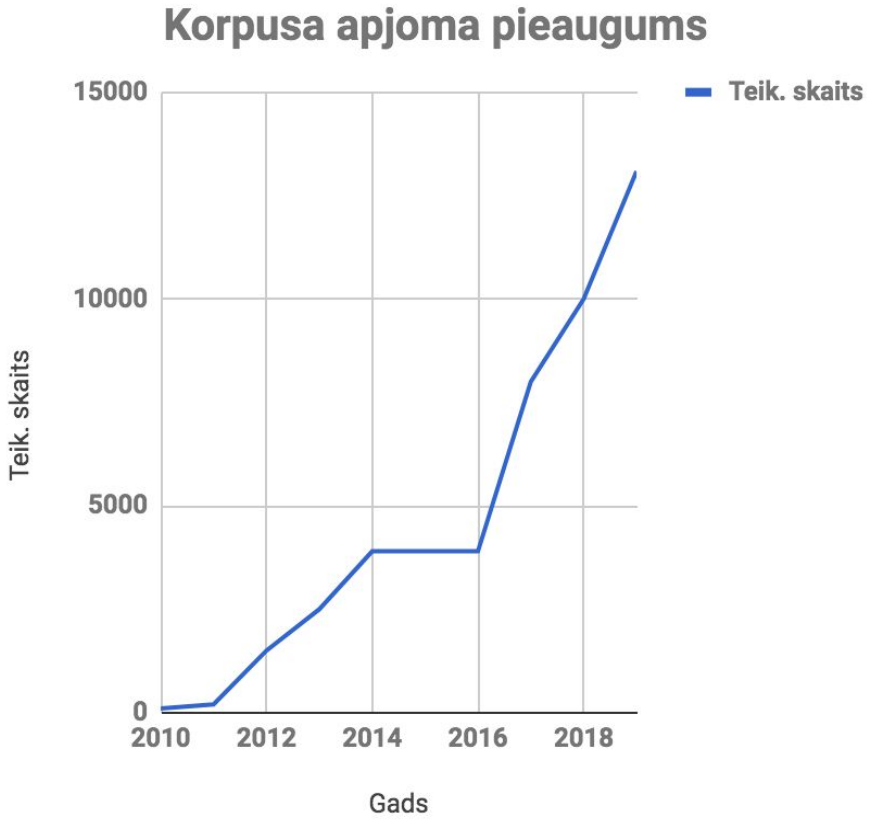
- statistikas vākšana

  Cik bieži "būt" lietots patstāvīgā funkcijā un cik – palīgverba funckijā?

Korpusa izveide

Projekti/programmas

- 2007–2010 SemTi-kamols
- 2010–2012 VPP
- 2013–2014 LETA
- 2015–2016 ---
- 2017–2019 FullStack



Kas lācītim vēderā?

- Teksti no LVK2018 (13090 teikumu)
- Pilni teksti > rindkopas
- Pārstāvēti biežāk lietotie verbi
- Pārstāvēti dažādi valodas paveidi – publicistika, daiļliteratūra, zinātniskie teksti u.tml.

Mākslīgā intelekta laboratorija
LU MII

Korpusa formāts

Sem-Ti-Kamols gramatikas modeļa formāts:

- izveidots LU MII;
- tajā veikta manuāla un pusautomātiska korpusa marķēšana;
- gramatikas modelis tuvāks latviešu valodas sintakses teorijai.

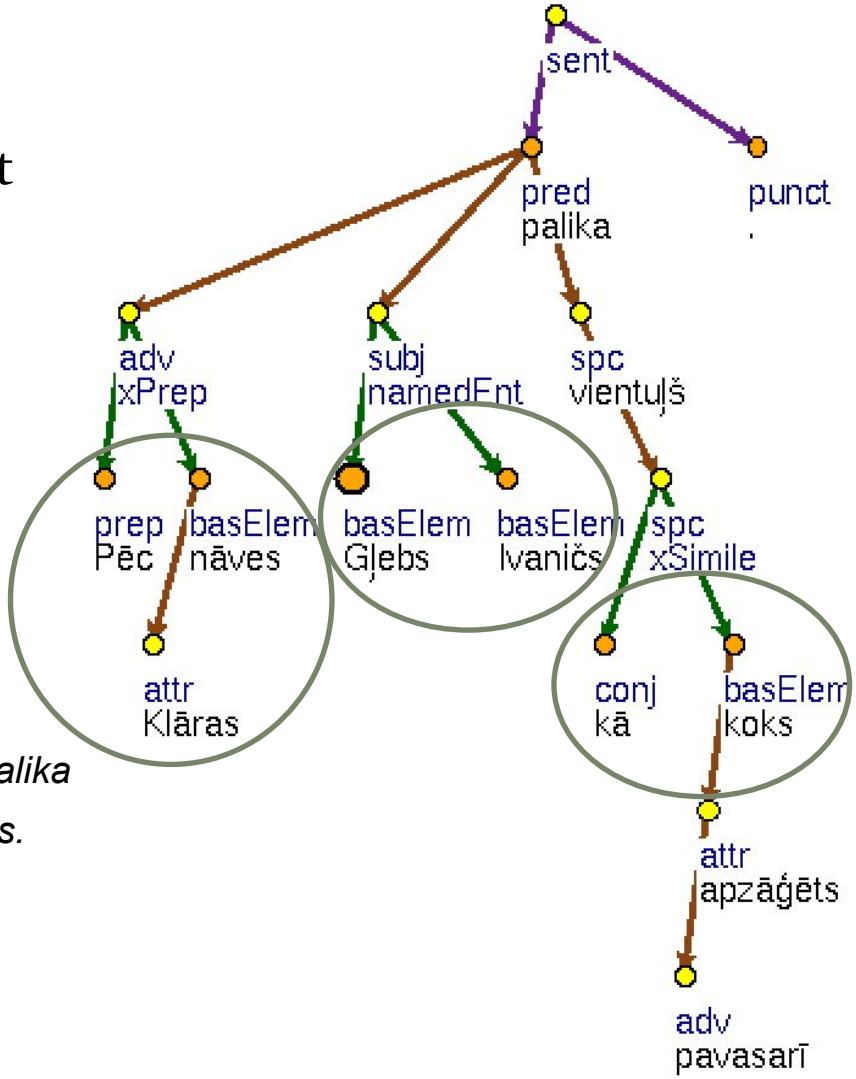
Universālo atkarību gramatikas modeļa formāts:

- starptautiska iniciatīva, iesaistīti vairāk nekā 100 korpusi 70 valodās;
- dati tiek automātiksi pārvērsti uz šo formātu;
- gramatikas modelis paredzēts, lai pēc vienota marķējuma varētu atrast līdzīgas valodas kategorijas;
- dažādu sintakses kategoriju izpratne atšķiras no LV tradīcijas.

Mākslīgā intelekta laboratorija
LU MII

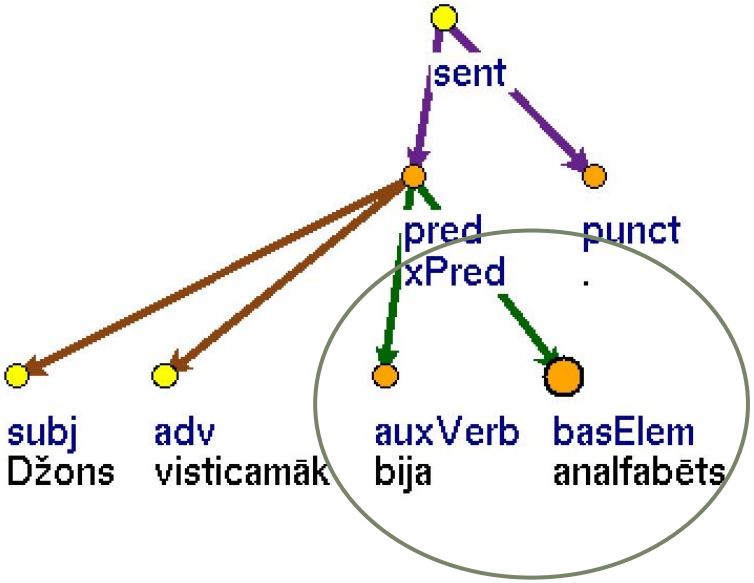
Hibrīds sintakses modelis

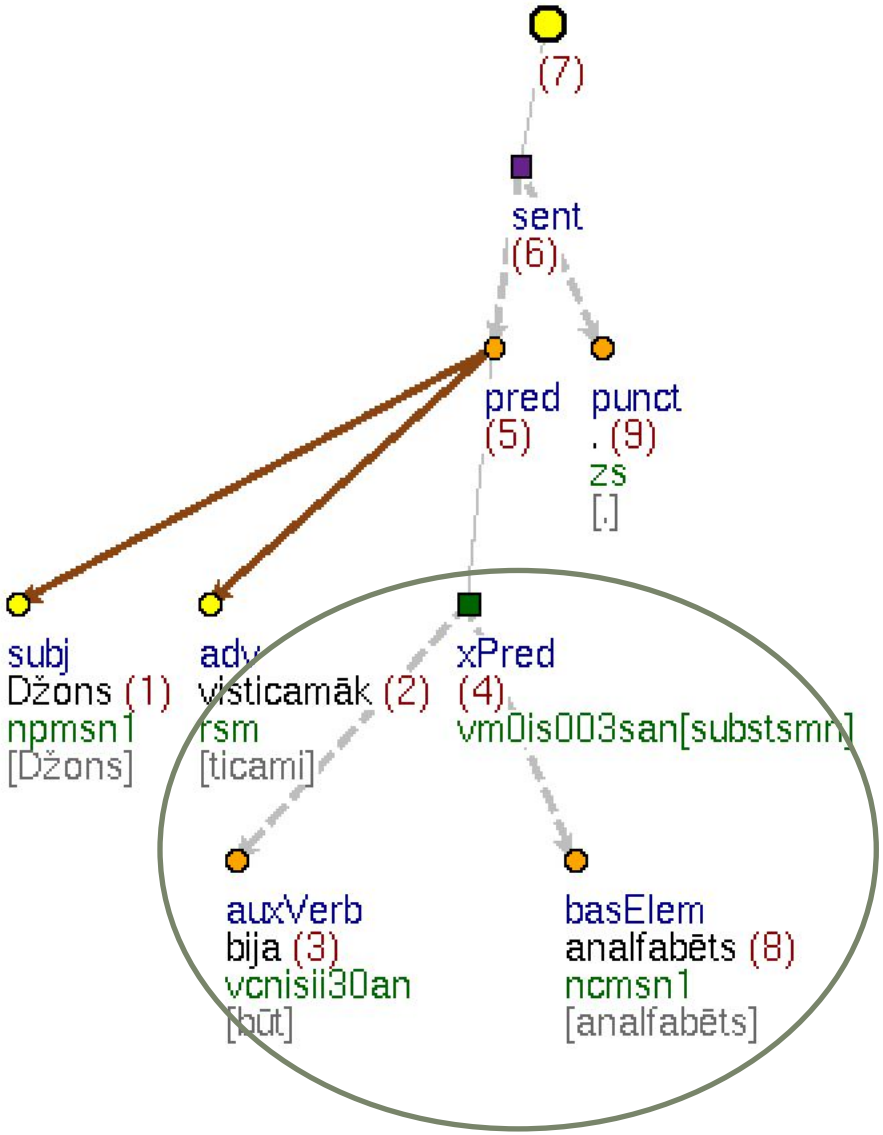
- **Atkarību gramatika** kombinēta ar **frāžu struktūras** gramatiku
- Sintakses modelis pēc iespējas atbilst latviešu valodas mūsdienu sintakses teorijai



Pēc Klāras nāves Gļebs Ivaničs palika vientuļš kā pavasarī apzāģēts koks.

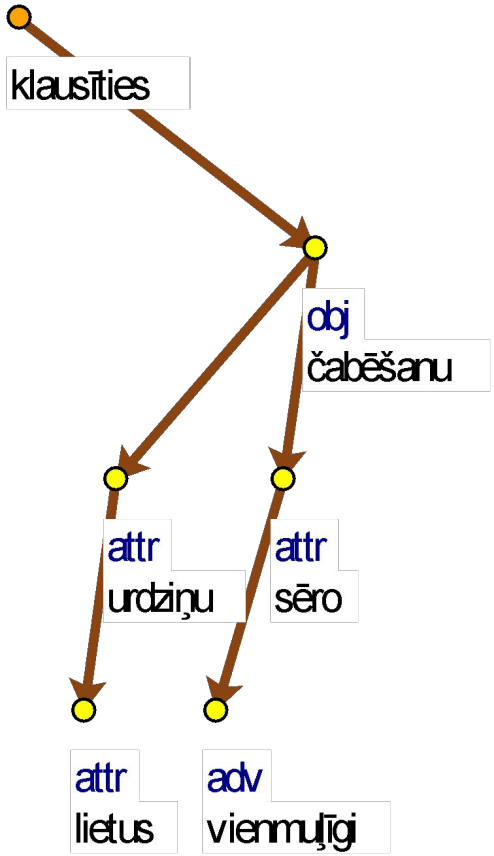
Teikums vienkāršotā un pilnā attēlojumā





Atkarību lomas

... klausīties lietus urdziņu vienmuļīgi sēro čabēšanu



Sintakses vienības	
Teikuma locekļi	teikuma priekšmets, izteicējs, apzīmētājs, u.c.
Ārpusshēmas komponenti	situants, determinants, sekundāri predikatīvs komponents
Palīgvārdi	saikļi, partikulas
Palīgteikumi	apzīmētāja, papildinātāja, pamatojuma palīgteikumi u.c.
Īpašas konstrukcijas	iespraudumi, tiešā runa, uzruna

Frāzes

Frāzes sastāv no vairākām tekstvienībām, t.i., elementi, kas sintaktiskajos sakaros iesaistās kopā kā veselums un starp kuriem nav pakārtojuma:

- salikti teikuma locekļi (x-vārdi) - *bija aizgājis*
- sakārtoti teikuma locekļi vai teikuma daļas - *Jānis un Anna pastaigājās*
- pieturzīmju konstrukcijas - *Jānis, protams, neiebilda.*

Frāzes - salikti teikuma locekļi

X-vārdi	
xPred	salikts izteicējs
xPrep	prievārdiska konstrukcija
xNum	salikts skaitlis
xApp	pielikuma konstrukcija
xSimile	salīdzinājuma konstrukcijas daļa
xParticle	konstrukcija ar partikulu
xFunctor	vairākvārdu saikļi un partikulas
subrAnal	vārdkopas analogs
coordAnal	vārdrindas analogs
namedEnt	nosaukuma konstrukcija
phrasElem	frazeoloģisks elements
unstruct	nestrukturēta frāze

Frāzes - sakārtojums

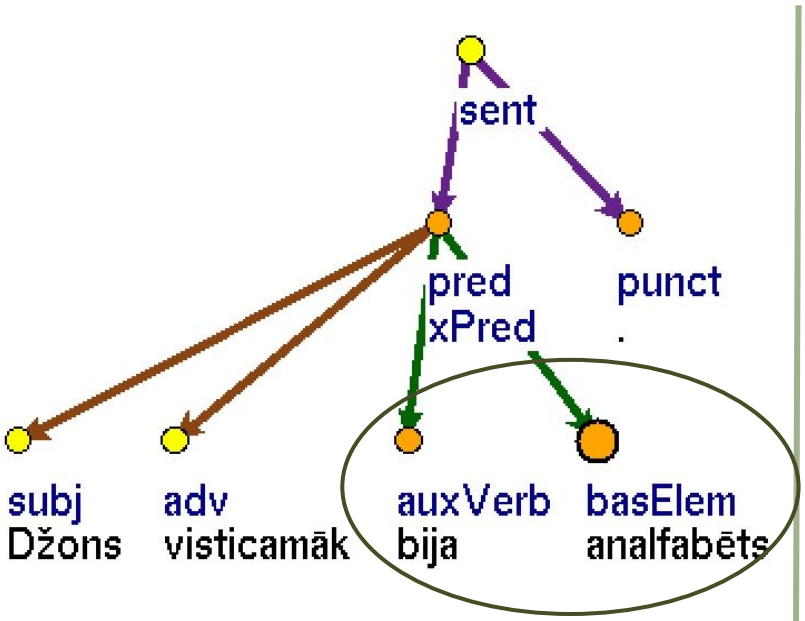
Sakārtojums	
crdParts	sakārtoti teikuma locekļi
crdClauses	sakārtotas teikuma daļas

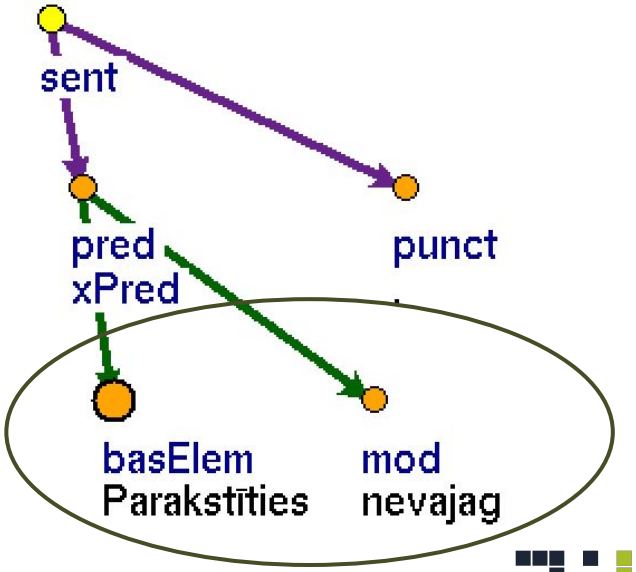
Frāzes - pieturzīmju konstrukcijas

Pieturzīmju konstrukcijas (PMC)	
sent	teikums
utter	izteikums
mainCl	salikta sakārtota teikuma daļa
subrCl	palīgteikums
spcPmc	sekundāri predikatīva komponenta pieturzīmes
insPmc	iesprauduma pieturzīmes
address	uzrunas konstrukcija
interj	izsauksmes vārdu pieturzīmes
dirSpPmc	tiešās runas pieturzīmes
qout	citāti un nosaukumi pēdiņās
particle	ar pieturzīmi atdalītas partikulas

1. Salikti izteicēji (xPred):

- saliktie laiki: *esmu darījis*
- sastata izteicēji: *bija analfabēts*
- izteicējs ar semantisko modificētāju: *parakstīties nevajag*

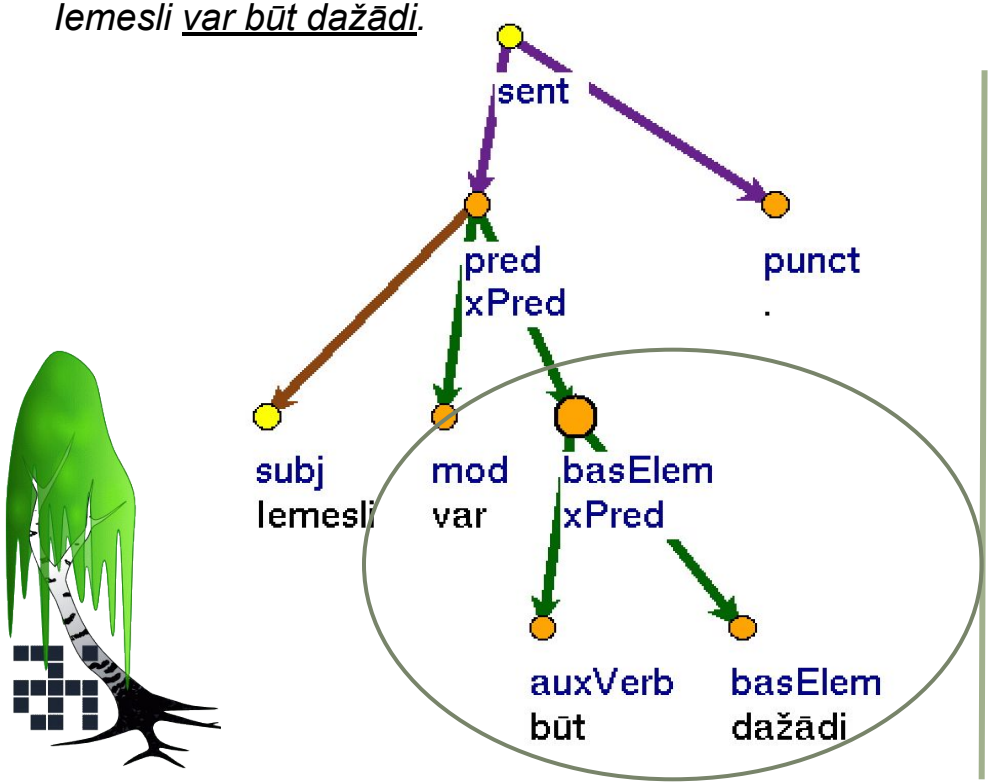




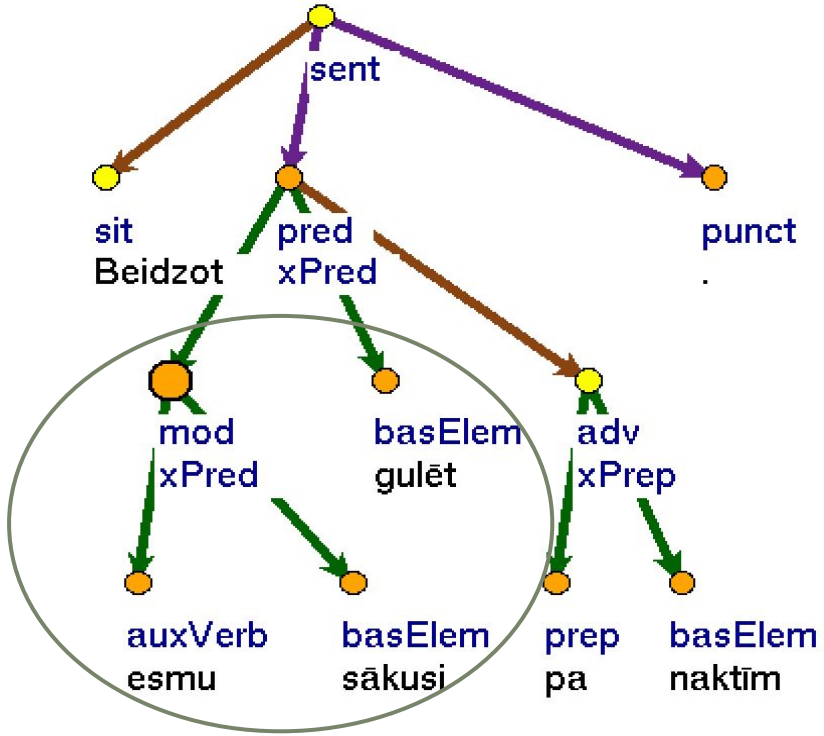
1.1. Vairākvārdu salikti izteicēji (xPred):

- sastata izteicējs ar modificētāju: *var būt dažādi*
- izteicējs ar saliktu modificētāju: *esmu sākusi gulēt*

Iemesli var būt dažādi.

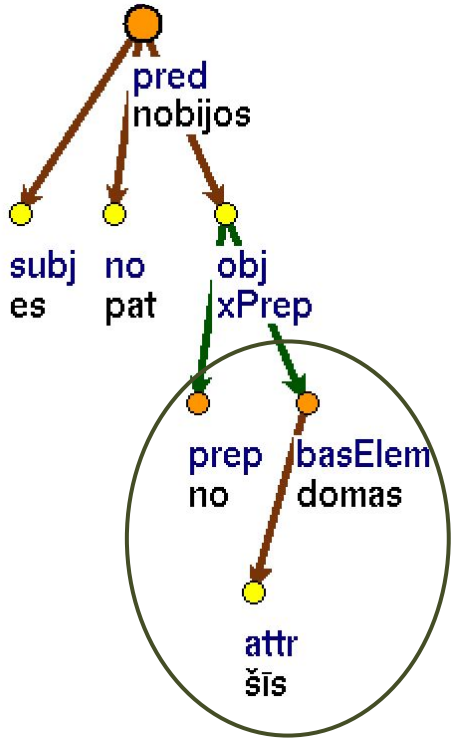


Beidzot esmu sākusi gulēt pa naktīm.



2. Prievārdiskas konstrukcijas (xPrep):

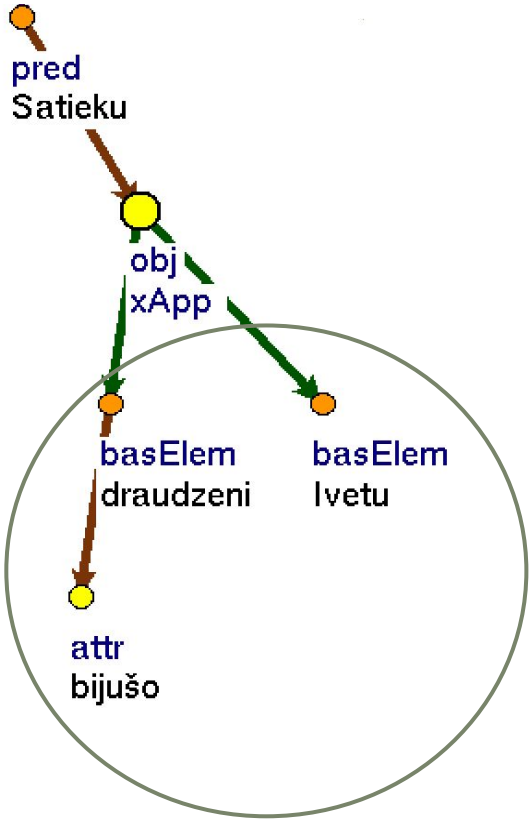
- ar prievārdu: *no domas, manis dēļ, uz skolu;*
- ar prievārdisku apstākļa vārdu: *durvīm blakus*



… es pat nobijos no šīs domas

3. Pielikuma konstrukcijas (xApp):

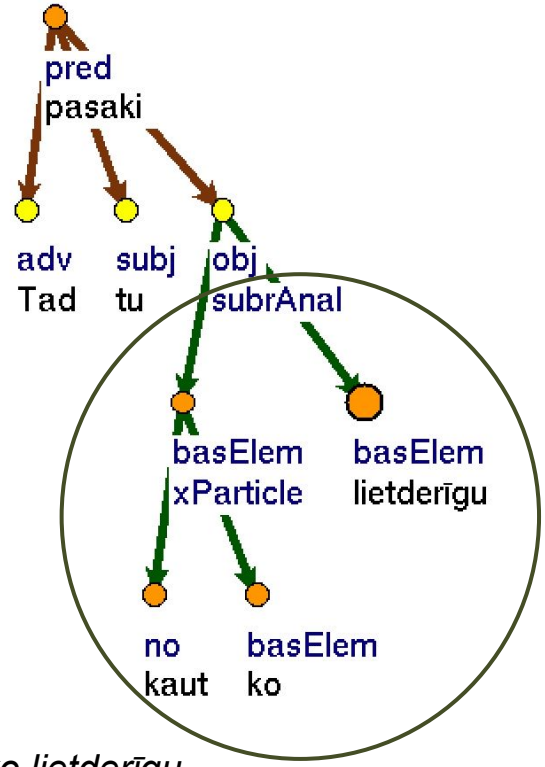
- saskaņots pielikums: *draudzene Iveta;*
- nesaskaņots pielikums: *laikraksts „Diena”.*



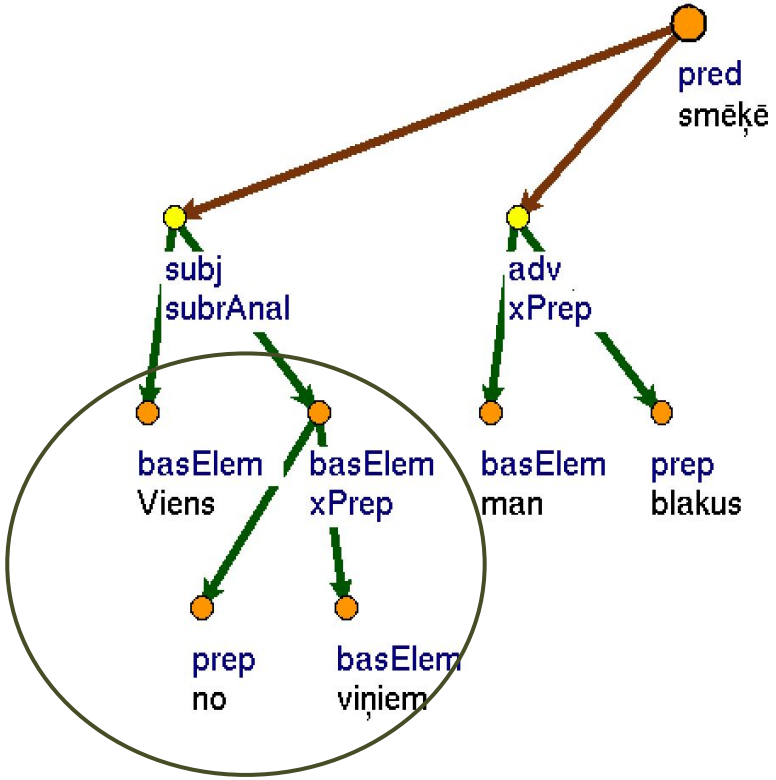
Satieku bijušo draudzeni Ivetu.

4. Vārdkopas analogs (subrAnal):

- vietniekvārdu savienojumi: *tas pats, mēs visi;*
- vietniekvārdu un īpašības vārdu savienojumi: *kaut kas lietderīgs, tāds priecīgs;*
- vietniekvārdu un skaitļa vārdu savienojumi: *abi divi, kādi pieci;*
- kopuma konstrukcijas: *viens no viņiem, kāds no klātesošajiem;*
- gramatizējušās salīdzinājuma konstrukcijas*: tāds kā noskumis.*



.. pasaki <u>kaut ko lietderīgu</u>.

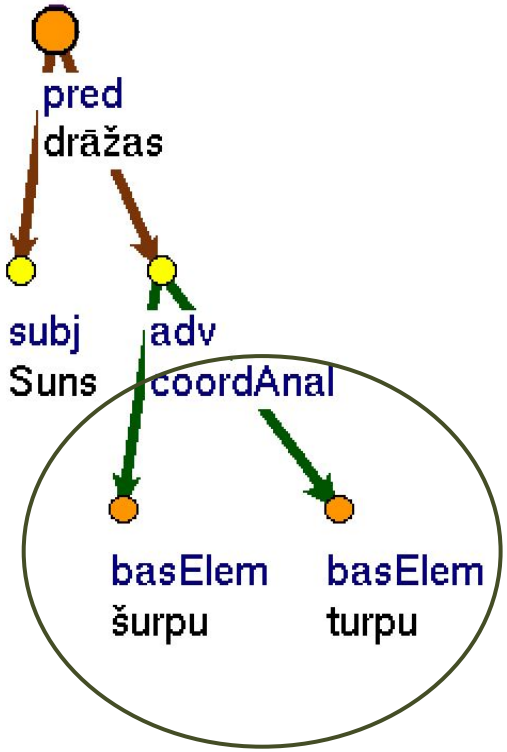


<u>Viens no viņiem</u> man blakus smēķē.

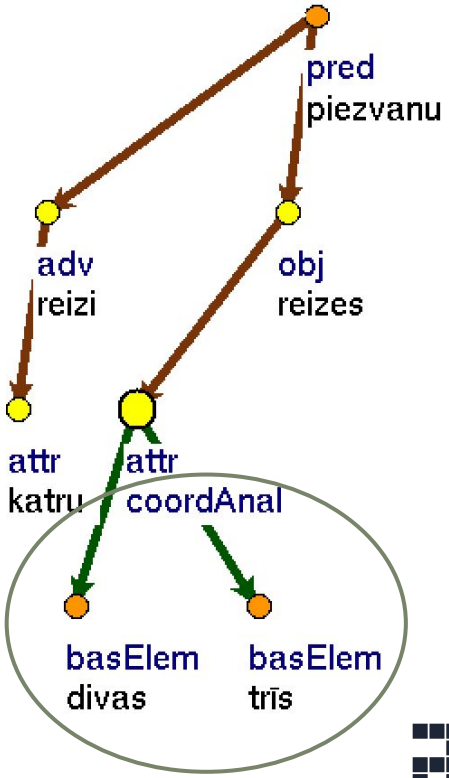
5. Vārdrindas analogs (coordAnal):

- *šurpu turpu; divas trīs*

Suns drāžas <u>šurpu turpu</u>...

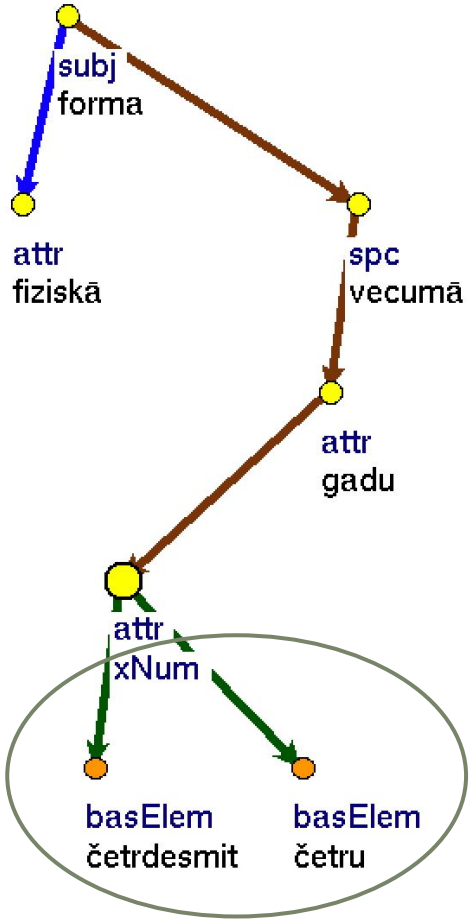


… katru reizi <u>divas trīs</u> reizes piezvanu…



6. Skaitļa vārdu savienojumi:

- *četrdesmit četri, simtu divdesmit trīs.*

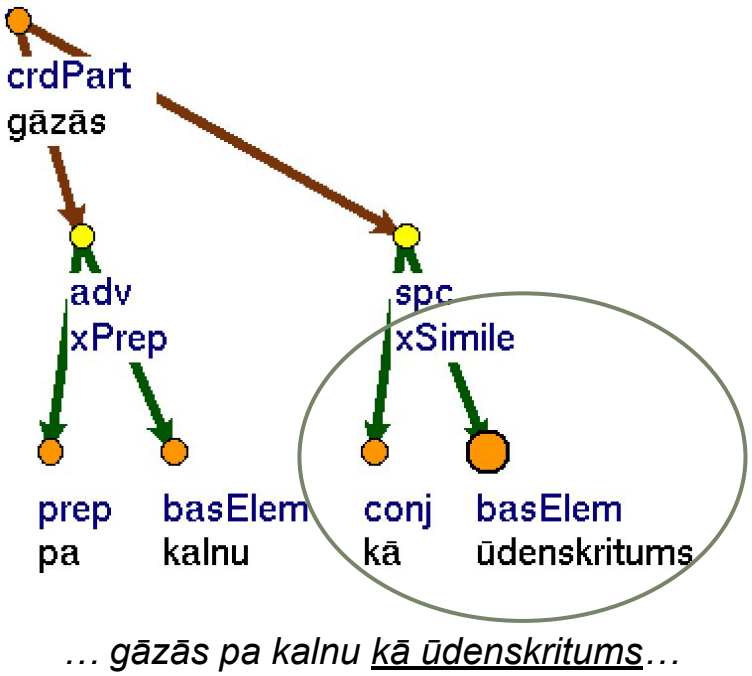


.. *četrdesmit četru gadu vecumā…*

Mākslīgā intelekta laboratorija
LU MII

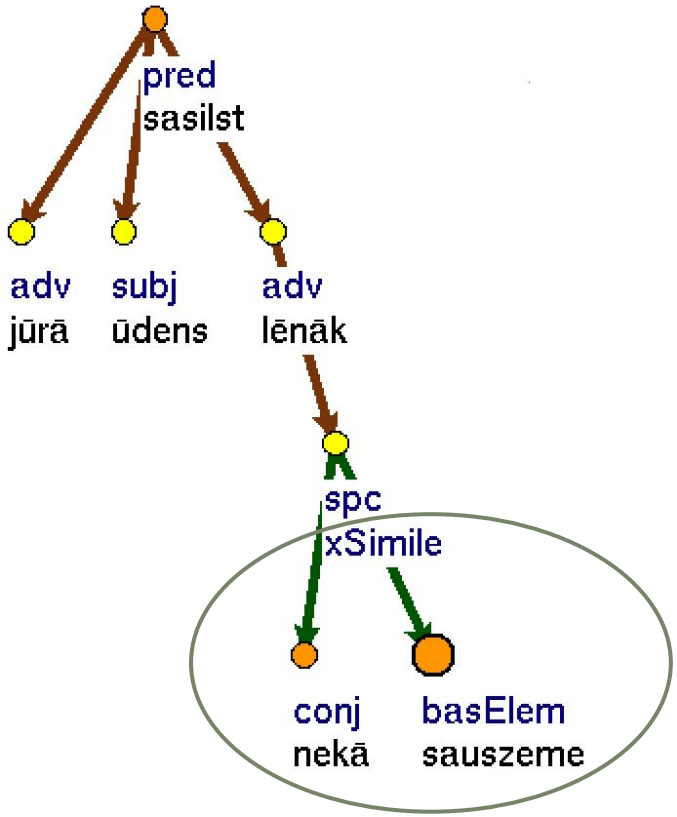
7. Salīdzinājuma konstrukcijas salīdzinātājdaļa (xSimile):

- pielīdzinājuma konstrukcijās: *kā ūdenskritums;*
- šķīruma konstrukcijās: *nekā sauszeme.*



… gāzās pa kalnu kā ūdenskritums…

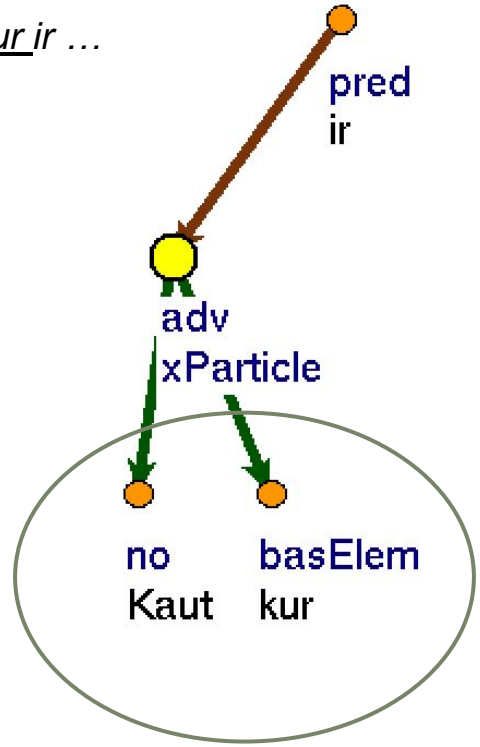
… jūrā ūdens sasilst lēnāk nekā sauszeme…



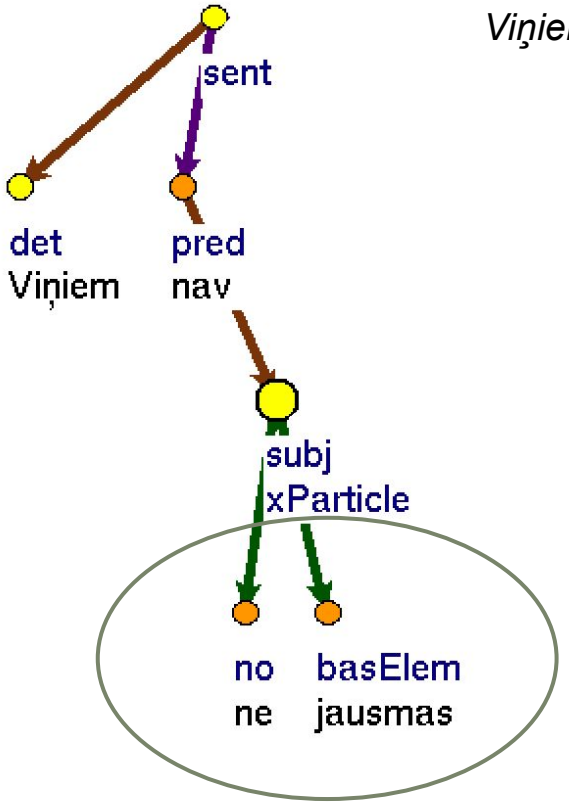
8. Vārdu savienojumi ar partikulu (xParticle):

- partikulas nozīmes apgabals: *kaut kur, nez kurš, it viss;*
- nolieguma apgabals*: ne vienmēr, ne reizi.*

Kaut kur ir …



Viņiem nav ne jausmas...

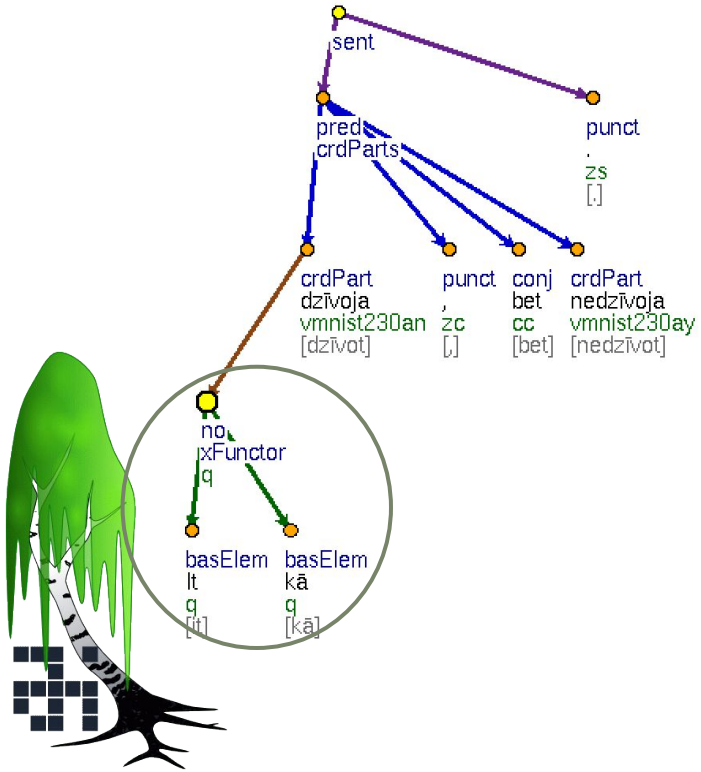


Mākslīgā intelekta laboratorija
LU MII

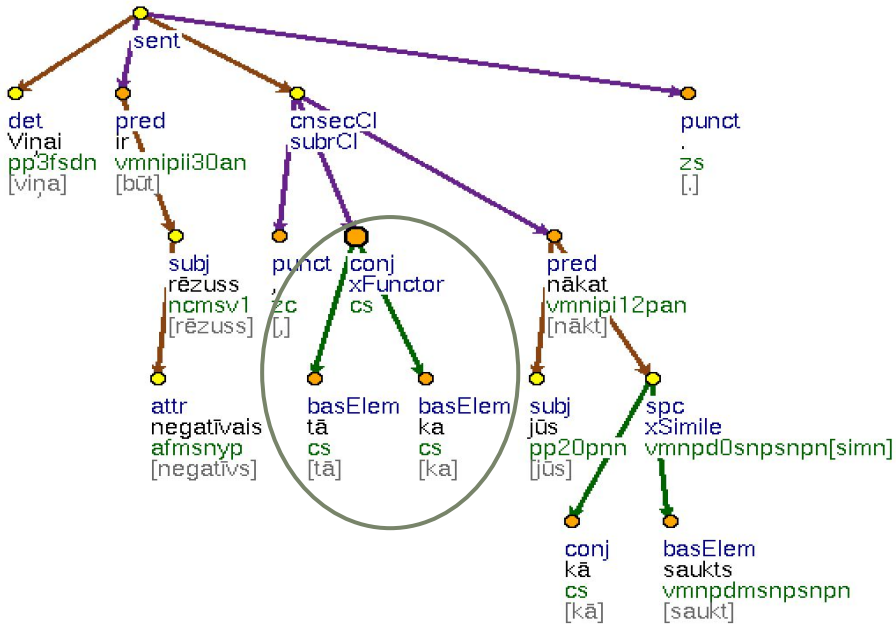
9. Vairākvārdu saikļi un partikulas (xFunctor):

- vairākvārdu saiklis: *vai arī, kaut gan, tā kā, tāpēc ka;*
- vairākvārdu partikula: *it kā, diez vai, gan jau.*

It kā dzīvoja, bet nedzīvoja.



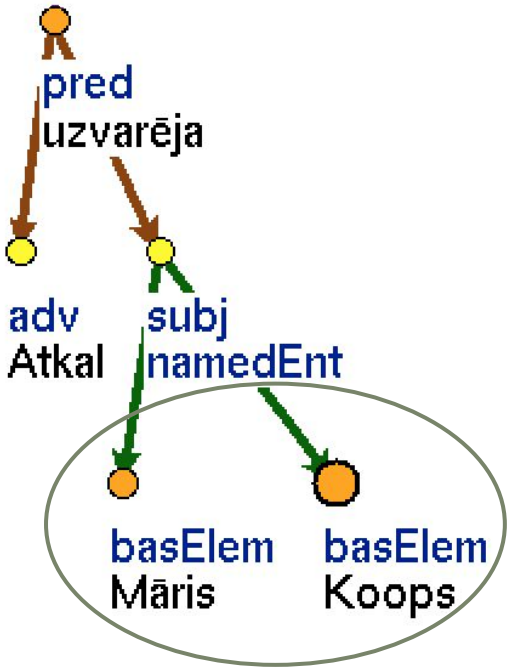
Viņai ir rēzuss negatīvs, tā ka jūs nākat kā saukts.



Mākslīgā intelekta laboratorija
LU MII

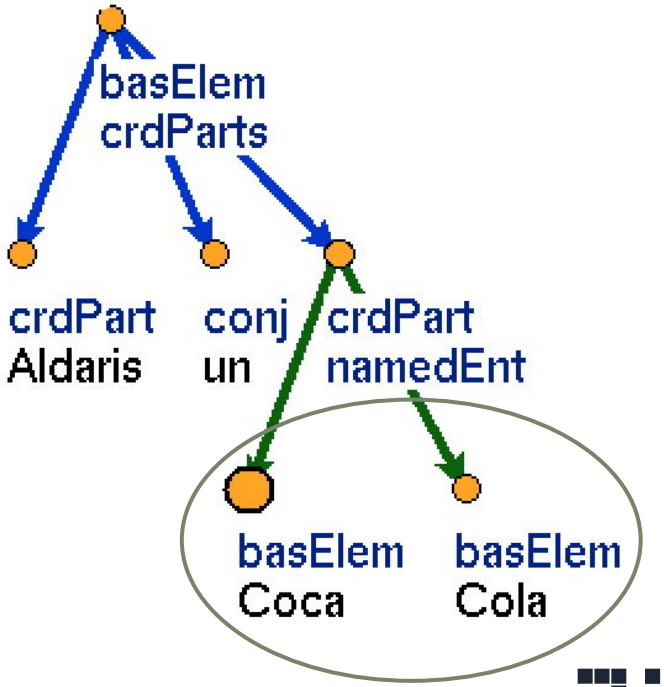
10. Nosaukumi (namedEntity)

- *Māris Koops, Bērziņš Investment, Coca Cola.*



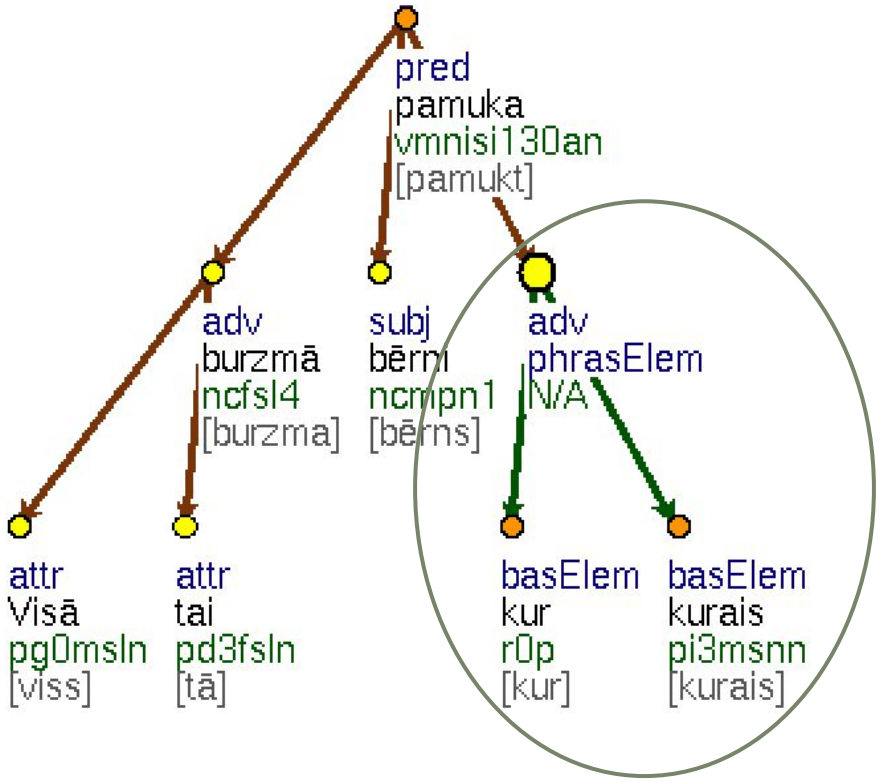
Atkal uzvarēja Māris Koops.

… Aldaris un Coca Cola…



11. Frazeoloģiskas vienības (phrasElem):

- *soli pa solim, viens un divi, ka tavu alma māteri, [bērni pamuka] kur kurais*

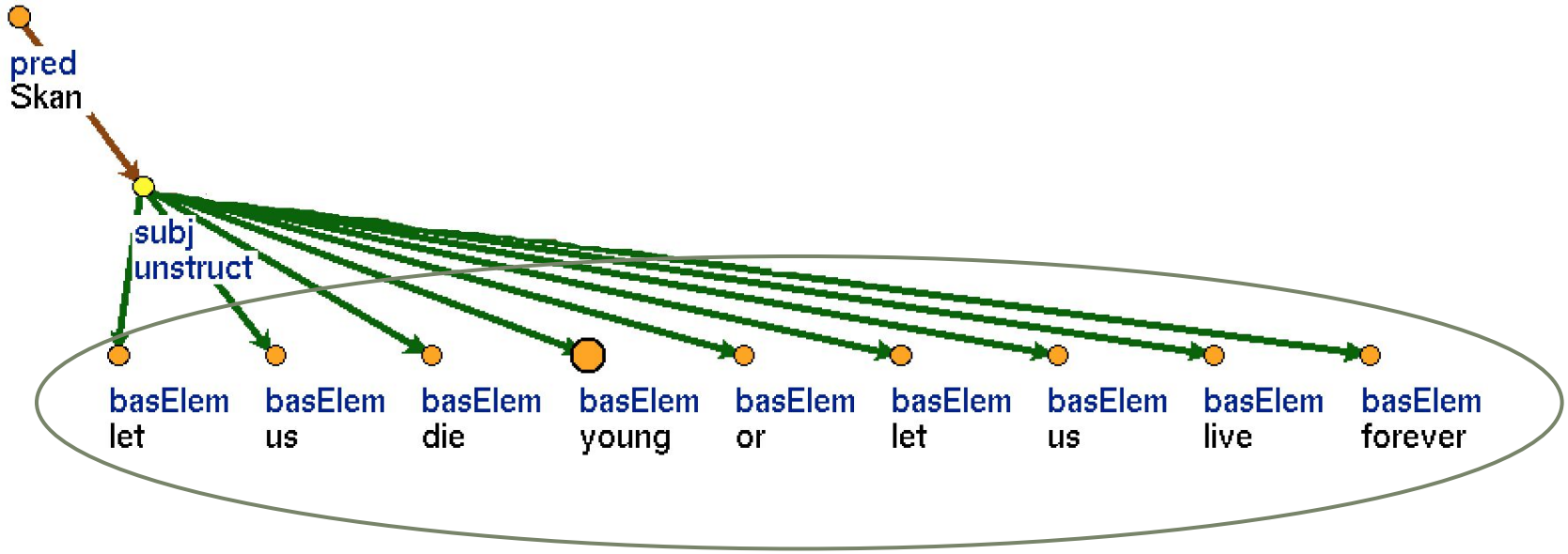


Visā tai burzmā bērni pamuka kur kurais ..

Mākslīgā intelekta laboratorija
LU MII

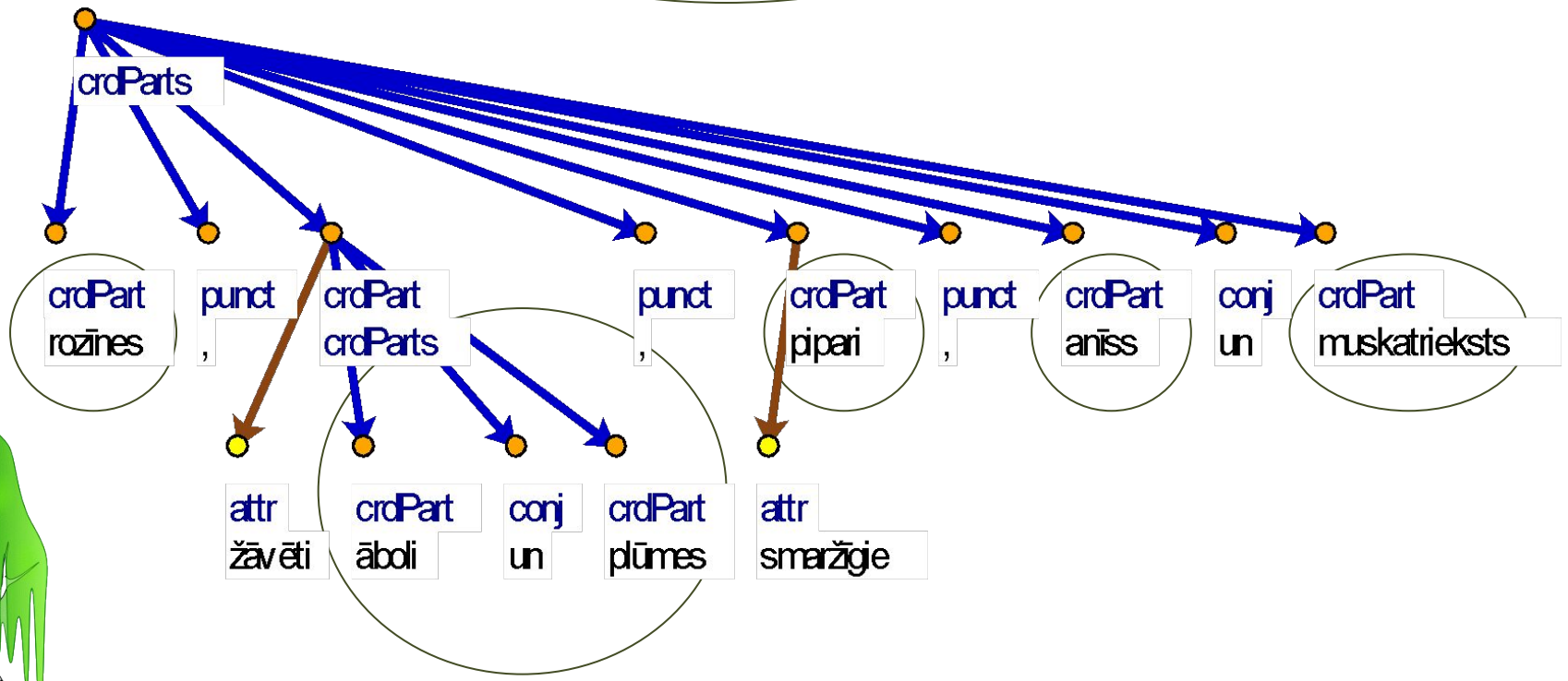
11. Nestrukturējamas frāzes (unstruct):

- formulas, izteicieni svešvalodā u.c.: *per aspera ad astra; 18x30*



Sakārtojums: sakārtoti teikuma locekļi

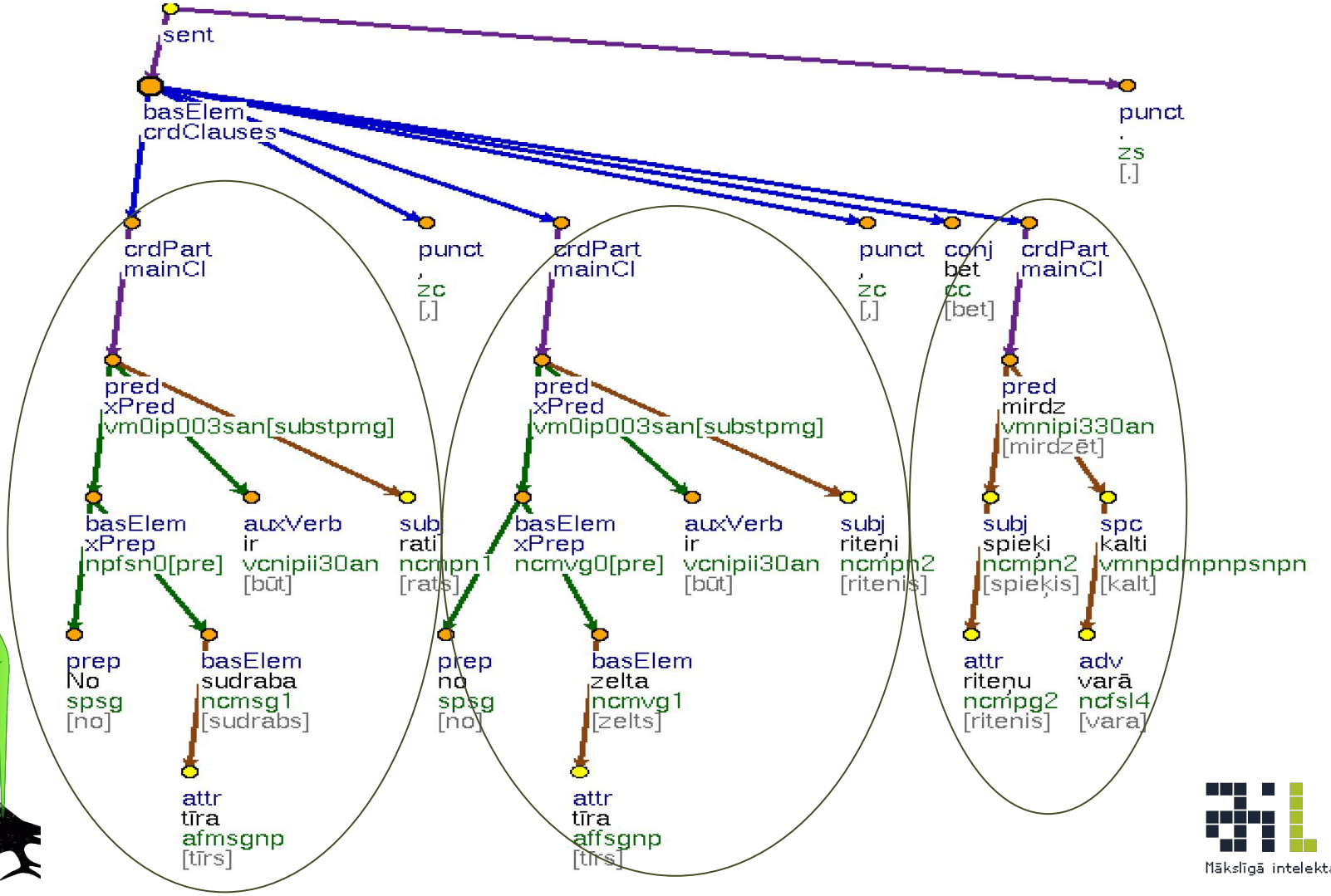
... rozīnes, žāvēti āboli un plūmes, smaržīgie pipari, anīss un muskatrieksts ...



Mākslīgā intelekta laboratorija
LU MII

Sakārtojums: sakārtotas teikuma daļas

No tīra sudraba ir rati, no tīra zelta ir riteņi, bet riteņu spieķi mirdz varā kalti.

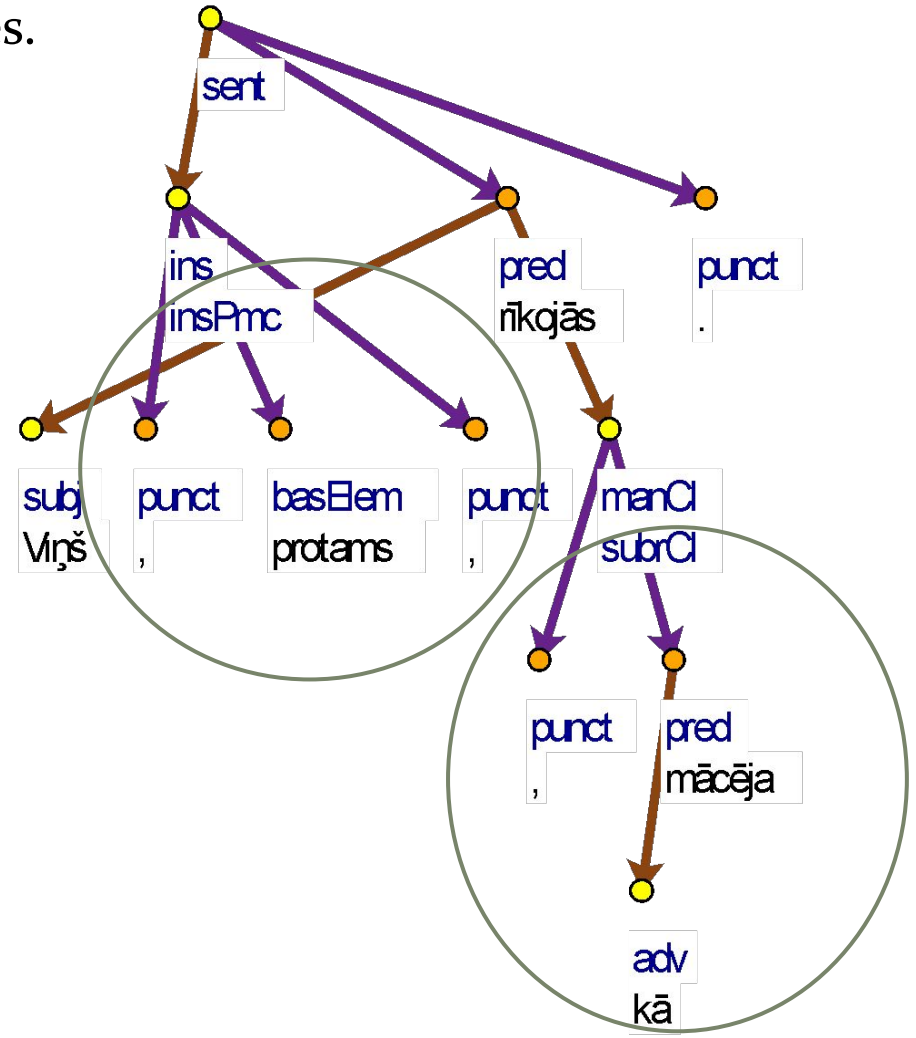


Mākslīgā intelekta laboratorija
LU MII

Pieturzīmju konstrukcijas

- Sastāv no teikuma elementa (daļas), kura dēļ pieturzīme(s) tiek lietota(s), un no pieturzīmes(-ēm). Ar tām teikuma struktūrā tiek iesaistītas pieturzīmes.

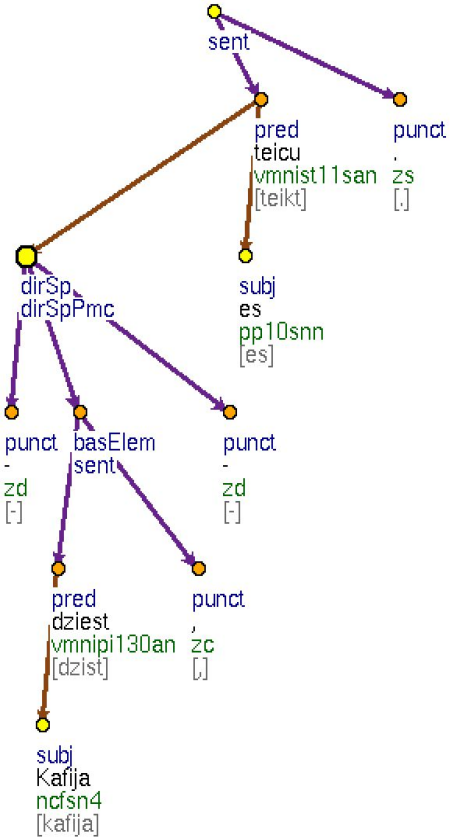
*Viņš, **protams,** rīkojās, **kā mācēja**.*

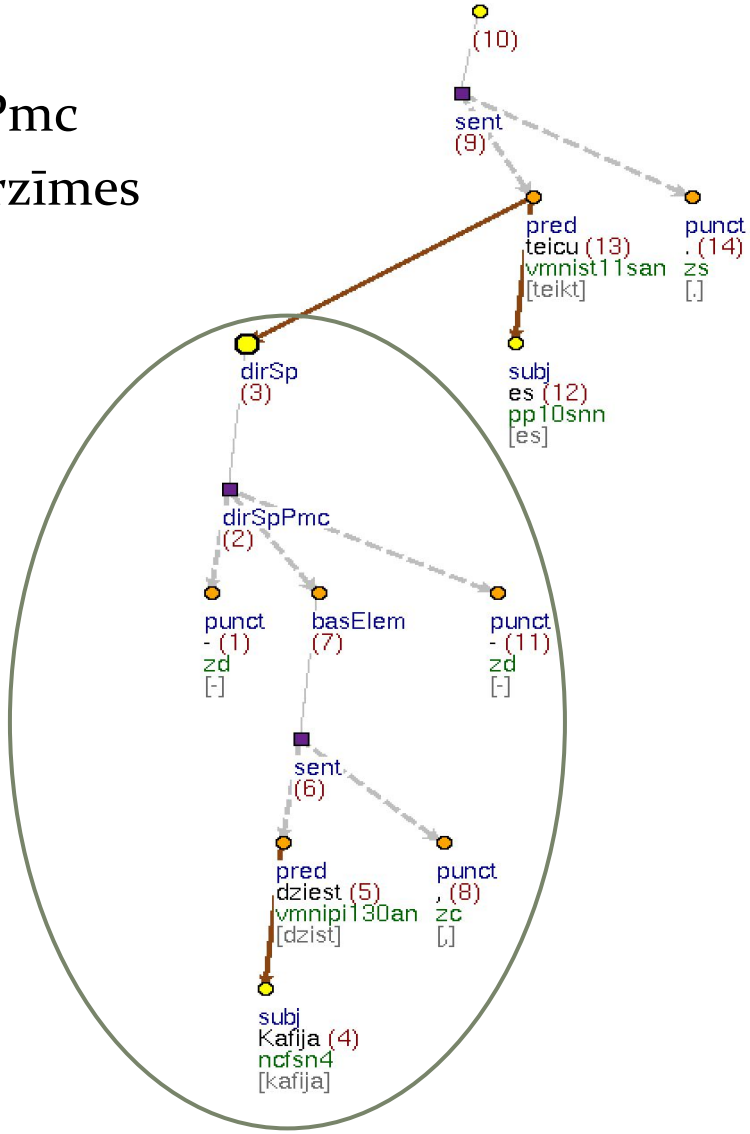


Mākslīgā intelekta laboratorija
LU MII

Pieturzīmju konstrukcijas - tiešā runa

- atkarību loma - dirSp
- konstrukcijas loma - dirSpPmc
- piesaista tiešās runas pieturzīmes





- Kafija dziest, - *es teicu.*

Mākslīgā intelekta laboratorija
LU MII

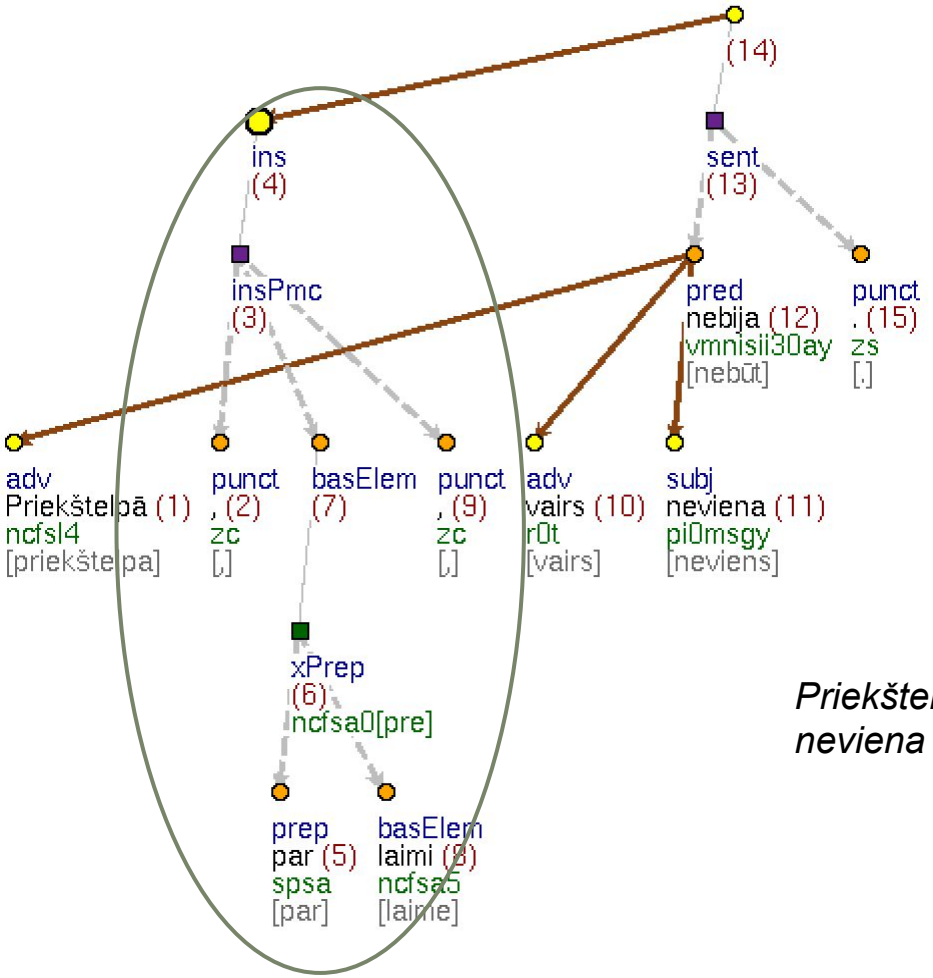
Pieturzīmju konstrukcijas - iespraudumi

- norādes uz informācijas avotu - *Ellasprāt, kā zināms*;
- norādes uz teksta autora attieksmi/vērtējumu - *protams, bez šaubām, šķiet, iespējams*;
- teksta iezīmētāji - norāda attieksmes starp teikuma elementiem vai teksta daļām - *tas ir, precīzāk, starp citu, citiem vārdiem sakot,*
- precizējumi un paskaidrojumi iekavās vai domuzīmēs - grāmatu autori, izdošanas gadi, lietas apraksts, u.c.

Mākslīgā intelekta laboratorija
LU MII

Pieturzīmju konstrukcijas - iespraudumi

- atkarību loma **ins**;
- konstrukcijas loma **insPmc;**
- var būt nepredikatīvas un predikatīvas vienības.



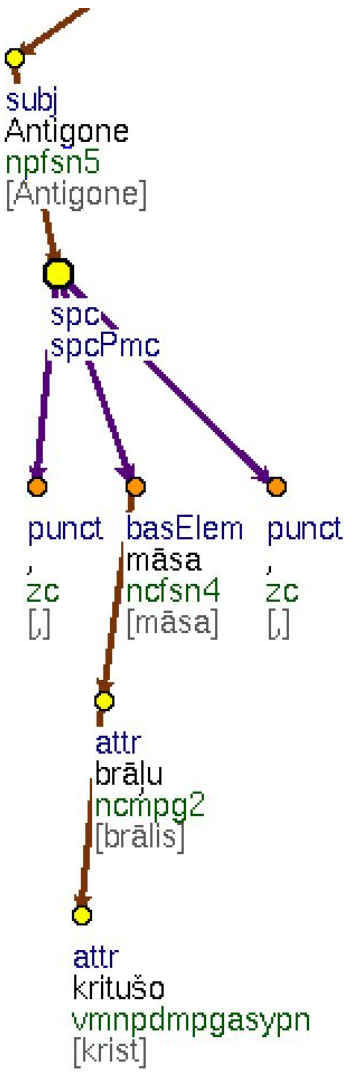
*Priekštelpā, **par laimi,** vairs neviena nebija.*

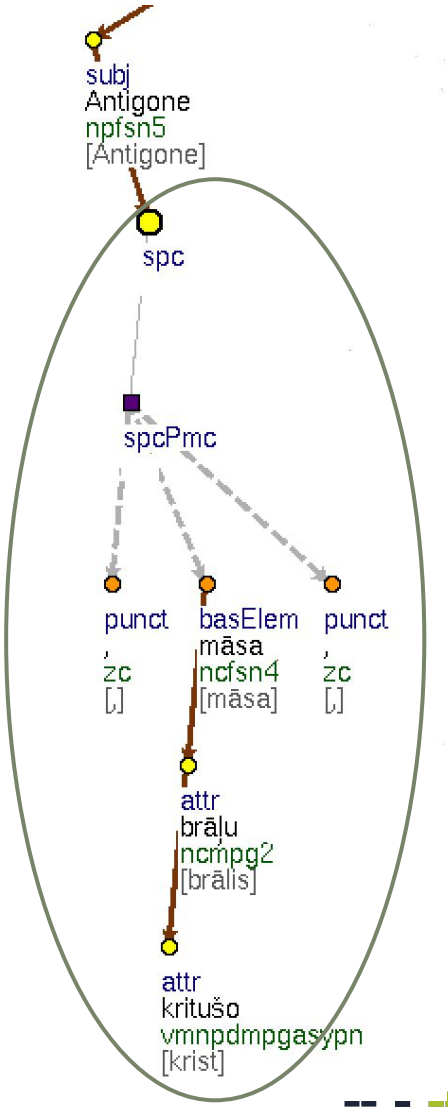
Mākslīgā intelekta laboratorija
LU MII

Pieturzīmju konstrukcijas - sekundāri predikatīvi komponenti

- divdabja teicieni;
- absolūtā datīva konstrukcija;
- savrupinājumi;
- paskaidrojošas vārdu grupas
- paralēli teikuma locekļi.

Antigone, ***kritušo brāļu māsa,*** ..





Hibrīdais sintakses modelis

Ieguvumi:

- palīdz attēlot latviešu sintakses teorijā aplūkotos saliktos teikuma locekļus;
- palīdz šķirt visas frāzes atkarīgos no frāzes elementa atkarīgajiem locekļiem;
- pieļauj datu transformāciju uz dažādiem atkarību modeļiem, nemainot marķēšanas principus un datus.

Sarežģījumi:

- datorlingvistikā nav attīstīti automātiskās sintaktiskās marķēšanas rīki šādam hibrīda modelim.

Turpmākais darbs:

- Samazināt manuāli pieļautu kļūdu un nekonsekventa marķējuma gadījumu skaitu;
- attīstīt gramatikas modeli - izveidot sīkāku sekundāri predikatīvo komponentu šķirumu;
- veicināt korpusa izmantošanu.

Publikācijas:

- Pretkalniņa L., Rituma L., Saulīte B. Deriving enhanced Universal Dependencies from a hybrid dependency-constituency treebank // Proceedings of the 21sh International Conference "Text, Speech, and Dialogue" (TSD), LNCS, Vol. 11107, Springer Link, 2018, pp. 95-105
- Gruzitis N., Pretkalnina L., Saulite B., Rituma L., Nespore-Berzkalne G., Znotins A., Paikens P. Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU // Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 2018, pp. 4506-4513
- Pretkalniņa L., Rituma L., Saulīte B. 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, Vol. 289, IOS Press, 2016, pp. 136–143
- Pretkalniņa L., Rituma L. Statistical syntactic parsing for Latvian. // Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), eds. Oepen S., Hagen K., Johannessen J.B., NEALT Proceedings Series, Vol. 16, ISBN 978-91-7519-589-6, Linköping University Electronic Press, Sweden, 2013, pp. 279–289
- Pretkalniņa L., Rituma L. 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, Vol. 247, IOS Press, 2012, pp. 185–192

Publikācijas:

- Pretkalniņa L., Nešpore G., Levāne-Petrova K., and Saulīte B. 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, eds. Candel Mora M.Á., Carrió Pastor M., ISBN 9788469462256, 2011, pp. 119–127
- Pretkalniņa L., Nešpore G., Levāne-Petrova K., and Saulīte B. A Prague Markup Language Profile for the SemTi-Kamols Grammar Model. // Proceedings of the 18th Nordic Conference of Computational Linguistics, Riga, 2011, pp. 303–306
- Pretkalniņa L., Levāne-Petrova K. Preparatory Work for Latvian Treebank // Proceedings of International Conference CORPUS LINGUISTICS – 2011, St.Petersburg, Russia, 2011, pp. 53-58
- Nešpore G., Saulīte B., Bārzdiņš G., and Grūzītis N. 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, Vol. 219, IOS Press, 2010, pp. 233–240
- Bārzdiņš G., Grūzītis N., Nešpore G., and Saulīte B. 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, Tartu, 2007, pp. 13–20

Mākslīgā intelekta laboratorija
LU MII

Paldies!

http://www.korpuss.lv

http://sintakse.korpuss.lv

Mākslīgā intelekta laboratorija
LU MII