

**CLARIN-LT**



# **Lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS: ateities perspektyvos**

**Agnė Bielinskienė, Loic Boïzou, Jolanta  
Kovalevskaitė, Erika Rimkutė**

**Vytauto Didžiojo universiteto  
Kompiuterinės lingvistikos centras  
CLARIN-LT centras  
2016 m.**

# Sintaksinė analizė

- Lietuvių kalbos sintaksiškai anototas tekstynas ALKSNIS rengtas 2015–2016 m.
- Tekstynas suplanuotas kaip lietuvių kalbos sintaksinės analizės aukso standartas.

# Įrankiai

- 2012–2015 m. VDU KLC vykdant projektą „*Lietuvių kalbos sintaksinės-semantinės analizės sistema tekstynui, lietuviškam internetui ir viešojo sektoriaus taikymams*“ sukurtas lietuvių kalbos **sintaksinis analizatorius** (L. Božou ir F. Zamblera).
- Priklausomybės medžiai (angl. *dependency trees*) generuojami PML (*Prague Markup Language*) formatu.
- Vizualizuojami ir redaguojami naudojant *TrED* redaktorių (Prahos Karolio universitetas).

Automatiškai anotuoti sakiniai peržiūrėti ir pataisyti lingvistų.





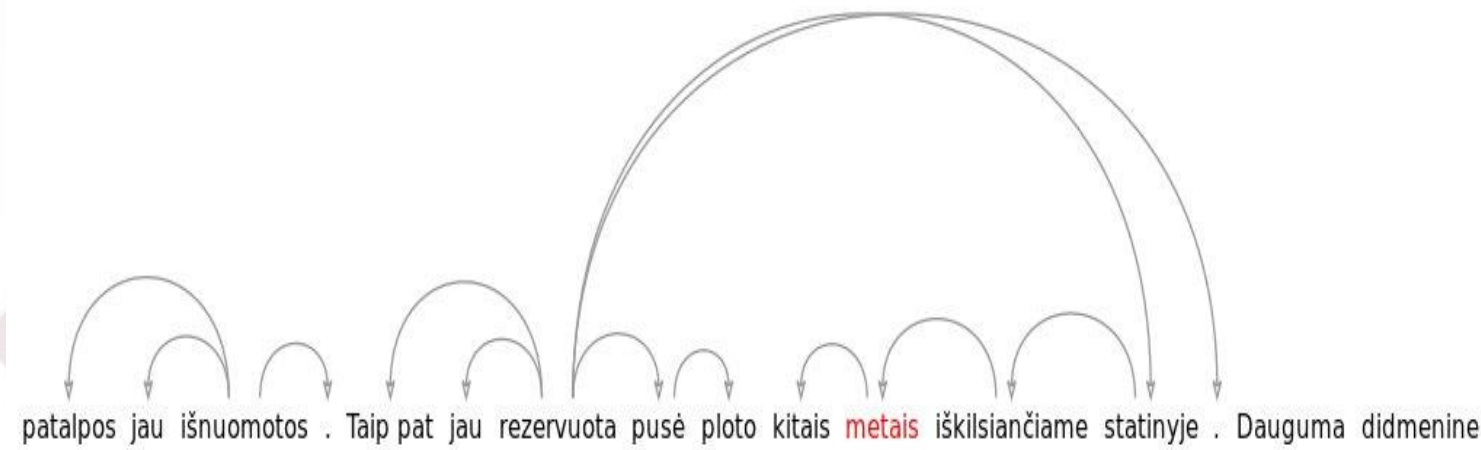
# Taip pat jau rezervuota pusė ploto kitais metais iškilsiančiame statinyje (PAULA)

1 ⓘ Path: Alksnis0-2\_paula > tb1-2-V2 (tokens 80 - 90)

left context: 10 ▾ right context: 5

|                       |       |        |      |                       |             |             |             |             |                      |             |       |
|-----------------------|-------|--------|------|-----------------------|-------------|-------------|-------------|-------------|----------------------|-------------|-------|
| išnuomotos            | .     | Taip   | jau  | rezervuota            | pusė        | ploto       | kitais      | metais      | iškilsiančiame       | statinyje   | .     |
| .PST.PL.F.PASS..NOM.. |       | .POS~. |      | .PST.SG.F.PASS..NOM.. | .F.SG.NOM.. | .M.SG.GEN.. | .M.PL.INS.. | .M.PL.INS.. | .FUT.SG.M.ACT..LOC.. | .M.SG.LOC.. |       |
| išnuomoti             | .     | taip   | jau  | rezervuoti            | pusė        | plotas      | kitas       | metai       | iškilti              | statinys    | .     |
|                       |       | pat    |      |                       |             |             |             |             |                      |             |       |
| VERB                  | PUNCT | ADV    | PART | VERB                  | NOUN        | NOUN        | PRON        | NOUN        | VERB                 | NOUN        | PUNCT |
| PredN                 | AuxK  | Adj    | AuxZ | PredN                 | Sub         | Atr         | Atr         | Adj         | Atr                  | Adj         | AuxK  |

⊞ deps (default\_ns)



# Anotuotų tekstynų svarba

- Kuriant lietuvių kompiuterinės analizės programas ir įrankius, itin svarbi gramatinė analizė ir statistinė informacija apie lietuvių kalbos vartoseną.
- Sintaksiškai anotuotas tekstynas vienas svarbiausių išteklių kiekvienos kalbos kompiuterinei analizei.

# ALKSNIO sandara

Tekstyną sudaro:

- 30 599 žodžiai
- 2355 sintaksiškai anotuoti sakiniai

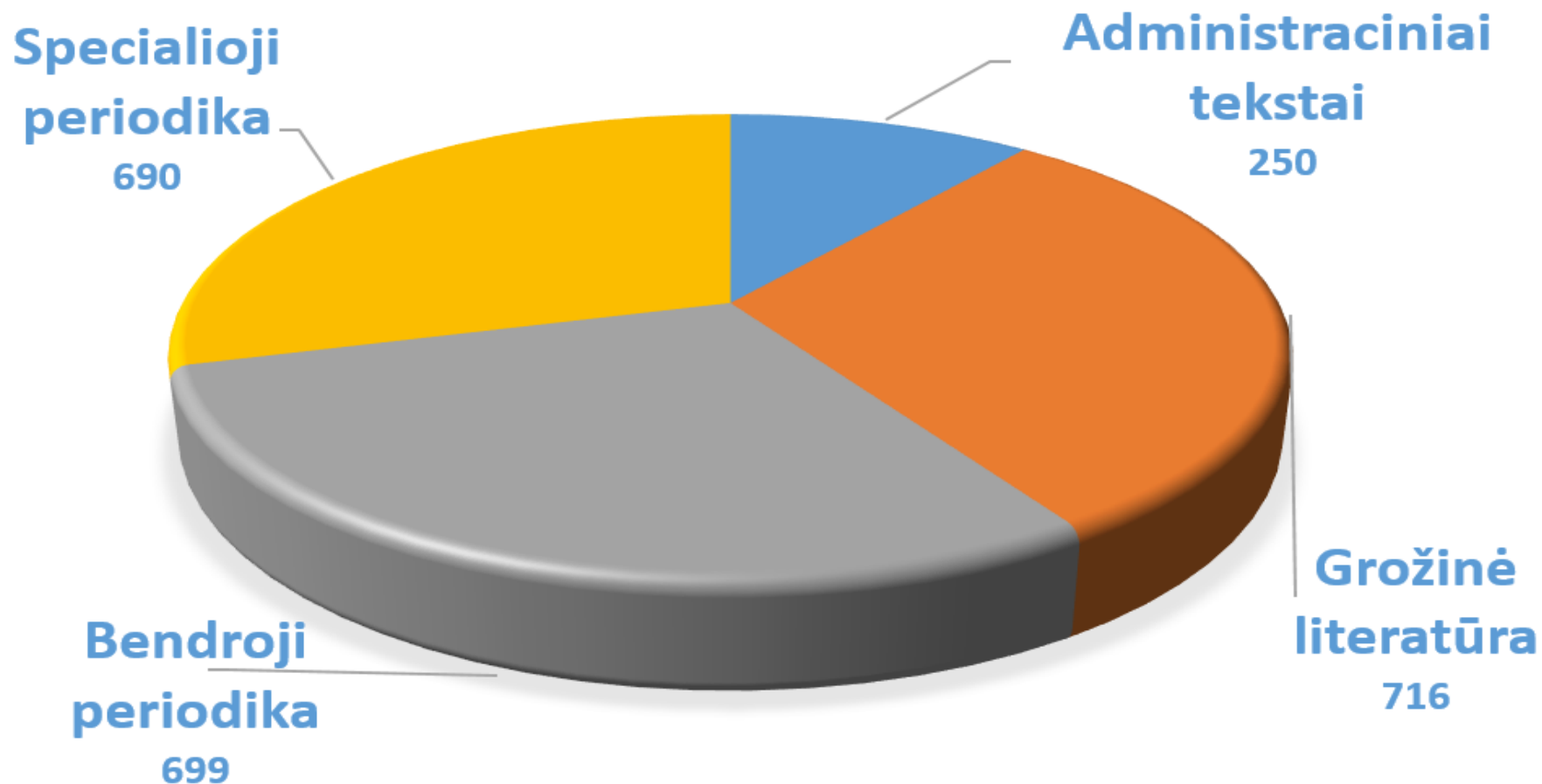
Tekstynas anotuotas automatiškai ir patikrintas lingvistų.

# Tekstyno dalys

Centre of  
Computational  
Linguistics

Kompiuterinės  
lingvistikos  
centras

## ALKSNIO DALIŲ SAKINIŲ SKAIČIUS



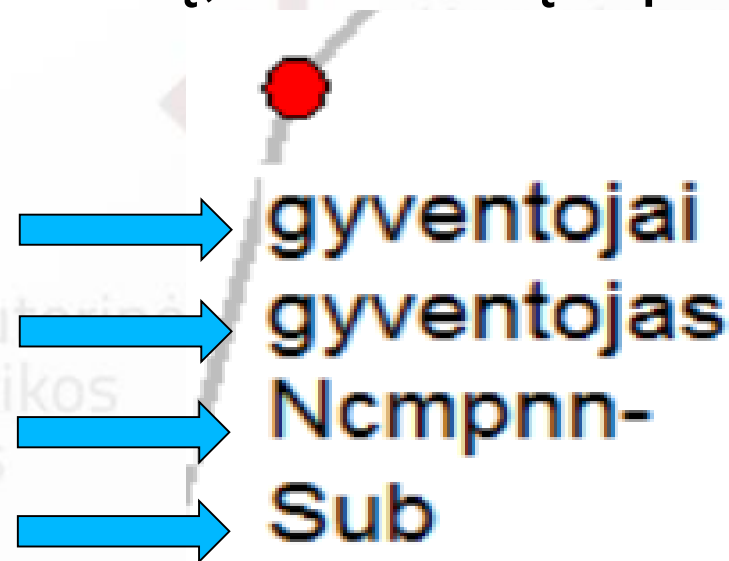


# Analizės lygmenys

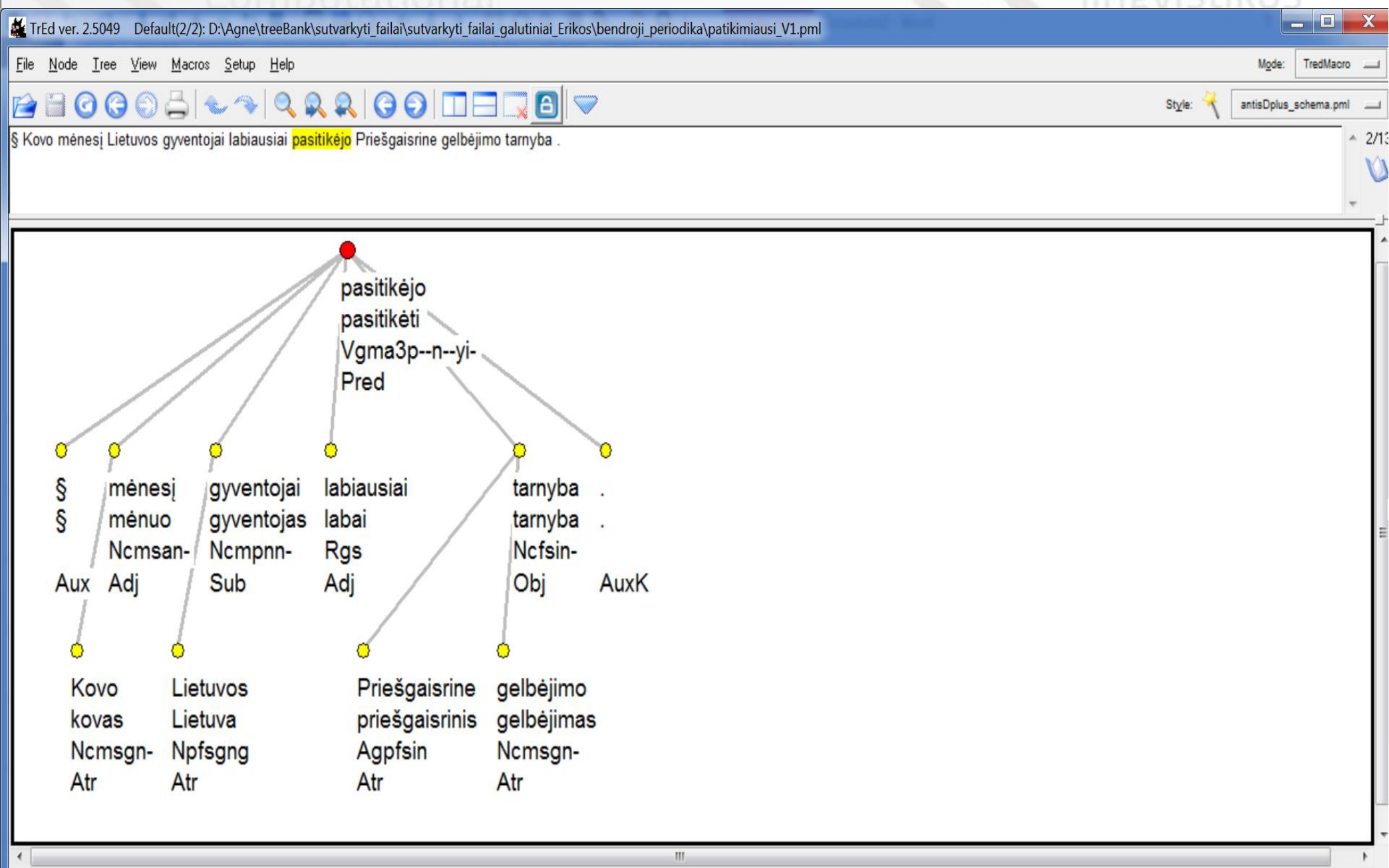
- Automatinis žodžių ir sakinių segmentavimas bei morfologinė analizė
- Automatinė sintaksinė analizė
- Rankinis sakinių tvarkymas

# ALKSNIO formatas

- Sakinys pateikiamas medžio principu (nurodomos priklausomybės).
- Kiekviena medžio viršūnė atitinka sakinio žodį, skyrybos ženklą ar kitą sakinio vienetą (simbolį, skaitmenį ir pan.).
- Prie visų žodžių nurodoma:
  - 1) konkreti žodžio forma,
  - 2) antraštinė forma (lema),
  - 3) morfologinės pažymos,
  - 4) sintaksinė funkcija.



# Sintaksiškai anotuotas sakinys

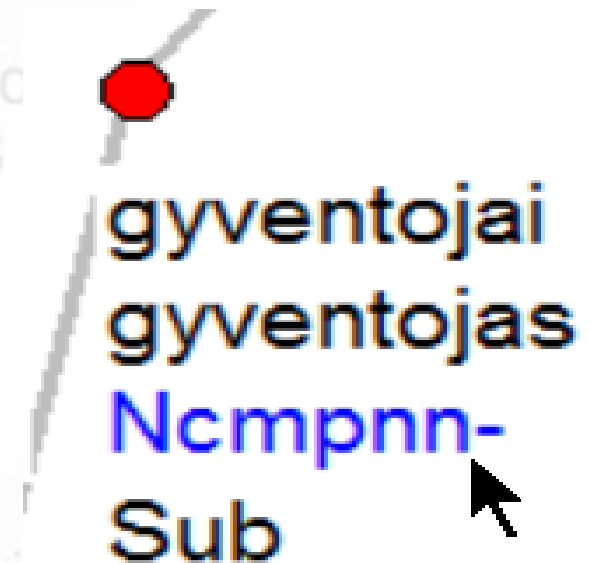


# Morfologinė analizė

- Morfologinės pažymos sudarytos remiantis *MULTEXT-East* formato pavyzdžiu.

- Naudojamas <http://semantika.lt/> prieinamas morfologinis anotatorius.

- Kiekviena kalbos dalis turi skirtingą morfologinių kategorijų skaičių (nuo 2 iki 14), kurios užrašomos santrumpomis.



gyventojai  
gyventojas  
Ncmpnn-  
Sub

# Morfologinis anotatorius ([www.semantika.lt](http://www.semantika.lt))

## Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema

### Lietuviško teksto anotavimas

[Paslaugos](#) / Lietuviško teksto anotavimas

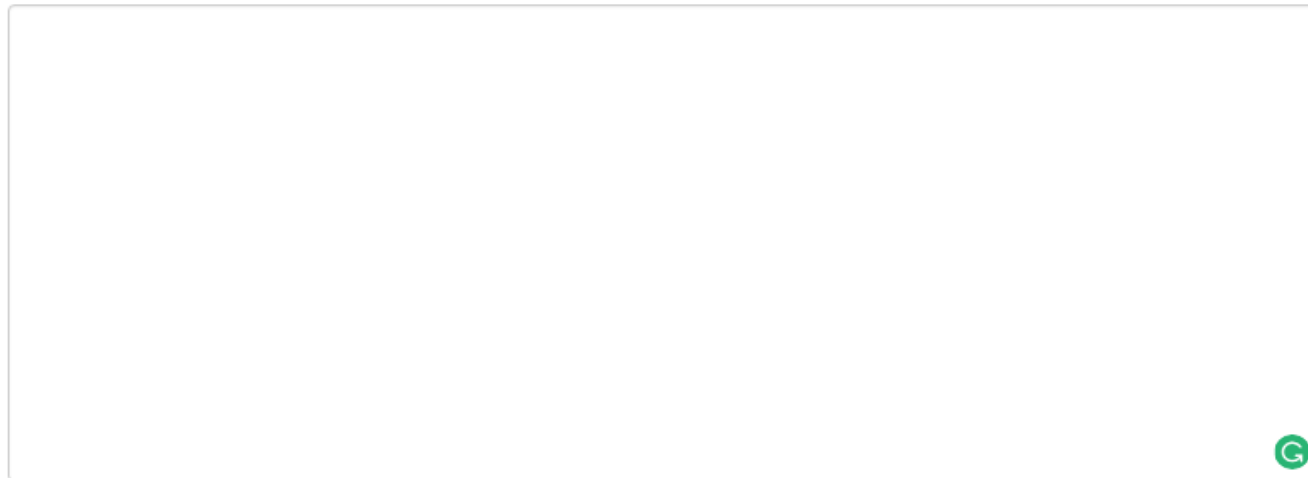
 [Paslaugos aprašymas](#)  [Pagalba](#)

#### Anotavimo parametrai

Grandinė\*

Morfologinė analizė ▼

Lietuvių kalbos tekstas\*



Anotuoti

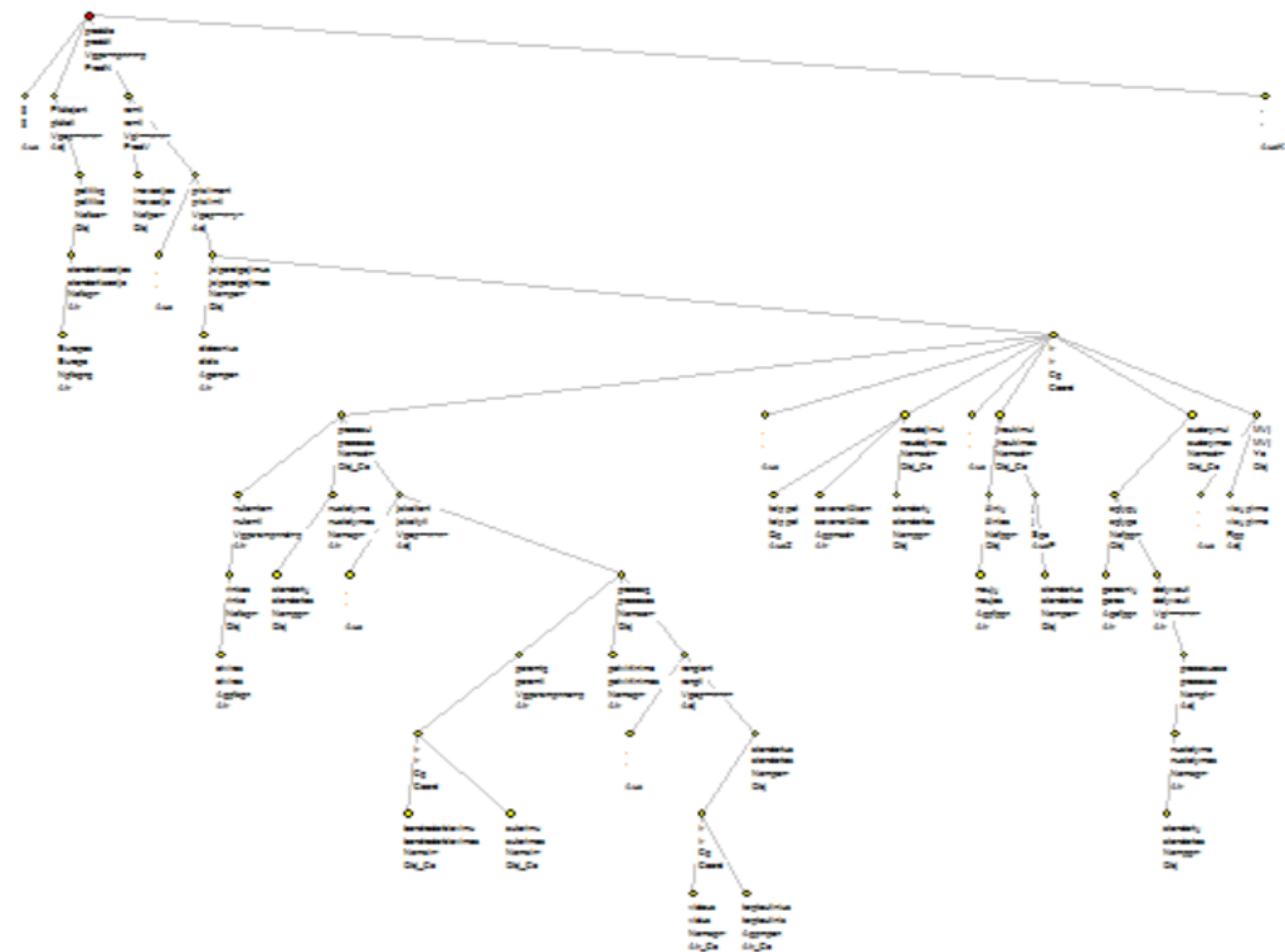
# Sintaksinė analizė

- Sintaksinis analizatorius:
  - apdoroja morfologiškai anotuotus failus ir pateikia sakinius, suskaidytus į dėmenis ar sintagmas;
  - generuoja sintaksinius medžius;
  - nurodo jų dėmenų sintaksines funkcijas.
- Remiamasi priklausomybių gramatikos (angl. *dependency grammar*) principais:
  - hierarchinė struktūra;
  - pagrindinis dėmuo – medžio viršūnė, kiti dėmenys prijungti remiantis priklausomybe.



## Computational





# Sintaksinės pažymos

**Pred (PredN, PredV)** – predikatas

**Sub** – subjektas

**Obj** – objektas

**Atr** – atributas (pažyminys)

**Adj** – aplinkybės

**Aux (AuxK, AuxC, AuxZ, AuxP, AuxL)** – tarnybinės k. d. arba kt.

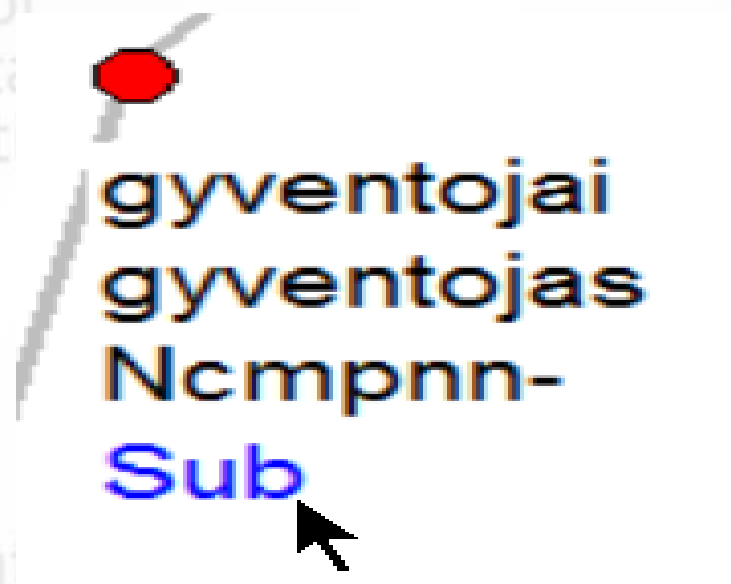
**Coord** – sujungimas

**Par** – įterpiniai

**ExD** – elipsė

**\_Co (Atr\_Co)** – sujungiamieji dėm.

**Pred\_ (Pred\_Obj)** – šalutinių sak. pažymos



# ALKSNIO pažymos ir sintaksinės funkcijos

| Pažyma                  | Sintaksinė funkcija                                                  | Pavyzdys                                                                                         |
|-------------------------|----------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| <b>Sub</b>              | Subjektas                                                            | <i>Jis (Sub) sakė</i>                                                                            |
| <b>Pred</b>             | Predikatas (arba pagalbinis žodis)                                   | <i>Jis ėjo (Pred)</i>                                                                            |
| <b>PredN</b>            | Daiktavardinė predikato dalis                                        | <i>Buvo (Pred) patenkintas (PredN)</i>                                                           |
| <b>PredV</b>            | Veiksmažodinė predikato dalis                                        | <i>Turi atsilyginti (PredV)</i>                                                                  |
| <b>Obj</b>              | Objektas                                                             | <i>Laukiu svečio (Obj); reikia spręsti (Obj)</i>                                                 |
| <b>Atr</b>              | Attributas (pažyminys)                                               | <i>Vidurinė (Atr) mokykla; šalis narė (Atr)</i>                                                  |
| <b>Adj</b>              | Aplinkybės                                                           | <i>Dabar (Adj) nuspręs; Partizanų gatvėje (Adj)</i>                                              |
| <b>Aux</b>              | Pagalbinė funkcija (e.g. kableliai ar kiti simboliai)                |                                                                                                  |
| <b>AuxC</b>             | Subordinacija (prijungiamieji sakiniai)                              | <i>Sakė, kad (AuxC)</i>                                                                          |
| <b>AuxK</b>             | Sakinio pabaigos skyrybos ženklas                                    |                                                                                                  |
| <b>AuxL</b>             | Pagalbinis leksinis vienetas (pvz., žodis užsienio kalba)            | <i>JAV kompanija North (AuxL) American (AuxL) Investment (AuxL) Consulting (AuxL) Inc. (Atr)</i> |
| <b>AuxP</b>             | Prielinksnis                                                         | <i>Pereis į (AuxP) lygį</i>                                                                      |
| <b>AuxZ</b>             | Dalelytė                                                             | <i>Kaip ir (AuxZ) visada</i>                                                                     |
| <b>Coord</b>            | Sujungimas                                                           | <i>Gamintojų ir (Coord) pardavėjų</i>                                                            |
| <b>_Co (Sub_Co)</b>     | Sujungiamieji dėmenys (pvz., vienas. s.d.)                           | <i>Gamintojų (Atr_Co) ir (Coord) pardavėjų (Atr_Co)</i>                                          |
| <b>Par</b>              | Įterpiniai, kreipiniai                                               | <i>Tiesą sakant (Par), atrodo</i>                                                                |
| <b>ExD</b>              | Elipsė ar kiti praleidimai sakinyje                                  | <i>Moteris – (Pred_ExD) būtybė</i>                                                               |
| <b>Pred_ (Pred_Atr)</b> | Prijungiamojo sakinio šalutinio dėmens tipas (žymima prie predikato) | <i>Rašė, kad ypatybės lemia (Pred_Obj)</i>                                                       |

# Ateities perspektyvos (1)

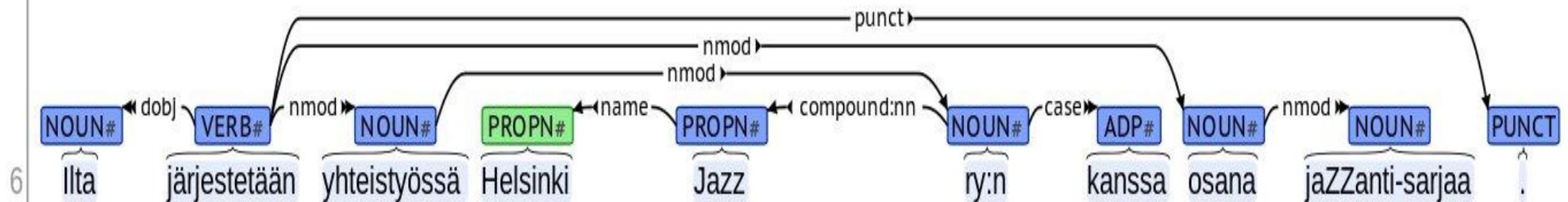
- Sintaksinių funkcijų tikslinimas, papildomų pažymų pridėjimas, anotavimo problemų sprendimas.
- Didinti sintaksiškai anotuotą tekstyną.
- Būtina pritaikyti *universal dependency* formatą (<http://universaldependencies.org/>).
- Sukurti statistiniu pagrindu veikiančią sintaksinę anotatorių.

# Ateities perspektyvos (2)

- Sutvarkyti teminių rolių (vaidmenų) anotavimą.
- Sužymėti pastoviuosius junginius (angl. *MWE*).
- Anaforų rezoliucija.
- Tekstyno prieinamumas vartotojams.



# *Universal Dependency* formatas



# Anotavimo problemų sprendimas

- Baigti rengti Sintaksinio anotavimo gaires.
- Peržiūrėti probleminius anotavimo atvejus:
  - Sudėtinių tarinių anotavimas
  - Elipsė
  - Bejungtukiai dėmenys ir aiškinimas
  - Objekto ir aplinkybių skyrimas
  - Tikslinamųjų aplinkybių skyrimas
  - Kai kurių kalbos dalių statusas
- Patikrinti iš naujo kai kuriuos anotavimo atvejus baigtame tekstyne.

# ALKSNIS

# CLARIN-LT



# saugykloje

ALKSNIS viešai prieinamas:

<http://hdl.handle.net/20.500.11821/10>

CLARIN-LT

NewsRepositoryLegal information

Talpykla / Įrašo peržiūra

Search

Lithuanian Treebank ALKSNIS

Please use the following text to cite this item or export to a predefined format:  
<http://hdl.handle.net/20.500.11821/10>

BIBTEXCMDI

CLARIN-LT

Authors

Rimkutė, Erika ; Bielinskienė, Agnė ; Kovalevskaitė, Jolanta ; Boizou, Loïc ; Aleksandravičiūtė, Gabrielė

Projekto URL

<https://clarin-lt.lt>

Problemos iškėlim...

2016-11-17

Tipas

corpus

Dydis

2355 sentences

Kalba(-os)

Lithuanian

Aprašymas

ALKSNIS v2

ALKSNIS v2 consists of 2,355 syntactically annotated sentences in the PML (Prague Mark-up Language) format. The format allows researchers to visualise and edit syntactic trees by the editor TrED (<https://ufal.mff.cuni.cz/tred/>).

Each node of a tree corresponds to a word, a punctuation mark or other text element (symbol, digit etc.) within a sentence. The following information is presented for each node: 1) a used form; 2) a lemma; 3) a morphology tag, and 4) a syntactic function (subject, object, etc.). Dependencies are shown by links between words.

The morphology tag set of the corpus is based on the MULTEXT-East format (<http://nl.ijs.si/ME/V4/msd/html/index.html>). Syntactically annotated sentences are corrected according to guidelines that were created by scientists of VMU CCL, following rules of Prague Dependency Treebank. All the sentences are being manually checked and corrected by a group of linguists.

The TrED editor and a style file is needed in order to view the files with .pml extension (with style file "styleData/alksnis.css")

CLARIN-LT

Ką galima daryti?

DEPOSIT

CITE

Peržiūra

> Elektroninė saugykla

Mano paskyra

Prisijungti

Statistika

Piwik statistika

BETA

Bendra informacija

Įkelti į saugyklą

Cituoti

Paraiškų gyvavimo ciklas

DUK

Apie

Pagalbos langas

**CLARIN-LT**



# **Lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS: ateities perspektyvos**

**Agnė Bielinskienė, Loic Boïzou, Jolanta  
Kovalevskaitė, Erika Rimkutė**

**Vytauto Didžiojo universiteto  
Kompiuterinės lingvistikos centras  
CLARIN-LT centras**

**V. Putvinskio g. 23, 216 kab., Kaunas  
klc@vdu.lt  
<http://clarin-lt.lt/>**