**AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, ERIKA RIMKUTĖ**

Vytauto Didžiojo universitetas

# LIETUVIŲ KALBOS MORFOLOGIŠKAI IR SINTAKSIŠKAI ANOTUOTI TEKSTYNAI

ESMINIAI ŽODŽIAI: tekstynas, automatinė morfologinė analizė, automatinė sintaksinė analizė, kalbos technologijos.

## ĮVADAS

Jau 25-erius metus Vytauto Didžiojo universiteto (VDU) Kompiuterinės lingvistikos centre (KLC)[1], kuriamos ir plėtojamos kalbos technologijos (Utka ir kt. 2017). VDU KLC parengti lietuvių kalbos ištekliai ir įrankiai yra pagrindas, be kurių neapsieinama kompiuterizuojant lietuvių kalbą ir pritaikant jai naujas technologijas. 2015 m. pradžioje Lietuva tapo visateise CLARIN ERIC nare, o 2015 m. liepos mėnesį VDU KLC ir VDU Informatikos fakultetas kartu su partneriais iš Vilniaus universiteto bei Kauno technologijos universiteto įkūrė CLARIN-LT konsorciumą ir pradėjo vykdyti projektą „Lietuvos narystė tarptautinėje mokslinių tyrimų infrastruktūroje – Bendroji kalbos išteklių ir technologijų infrastruktūra“[2], – taigi įsiliejo į bendrą Europos kalbų technologijų sistemą.

Šio straipsnio tikslas – pristatyti du lietuvių kalbos išteklius, anotuotus lietuvių kalbos tekstynus, parengtus VDU KLC: morfologiškai anotuotą tekstyną MATAS ir sintaksiškai anotuotą tekstyną ALKSNIS. Šie ištekliai viešai prieinami CLARIN-LT saugykloje, paiešką galima atlikti per ANNIS sistemą[3] (plačiau žr. 3 skyrių), taip pat paieška automatiškai morfologiškai anotuotame tekstyne galima svetainėje http://corpus.vdu.lt[4].

Anotuoti tekstynai – pagrindiniai ištekliai, atliekantys svarbų vaidmenį plėtojant kalbos technologijas. Kompiuterizuojant kalbą, rengiant kompiuterinius įrankius, skirtus automatinei kalbos analizei, būtina gramatinė analizė ir statistiniai duomenys apie kalbą. Taigi tekstynai, papildyti gramatinėmis pažymomis, yra tarsi žaliava tolesnei automatinei kalbos analizei. Tokie tekstynai panaudojami kitiems natūraliosios kalbos ištekliams ir įrankiams kurti tokiose srityse, kaip

[1] Prieiga internete: http://tekstynas.vdu.lt.
[2] Prieiga internete: http://clarin-lt.lt/.
[3] Prieiga internete: http://158.129.51.247:8080/annis-gui-3.4.4/.
[4] Tai iš 208 mln. žodžių sudarytas automatiškai morfologiškai anotuotas tekstynas. Toliau apie jį šiame straipsnyje nerašoma. Daugiausia dėmesio skiriama 1,6 mln. žodžių kalbininko peržiūrėtam morfologiškai anotuotam tekstynui MATAS.

automatinio kalbos atpažinimo sistemos, mašininis mokymas, automatizuotas vertimas, informacijos išgavimas ir pan.

Toliau straipsnyje atskirai pristatomi minėti tekstynai ir duomenų paieška juose per ANNIS sistemą. Tekstynais ir jų duomenimis gali pasinaudoti tyrėjai bei studentai, tiriantys lietuvių kalbos gramatikos sistemą, dėstytojai ir mokytojai, sudarydami užduotis, ir visi, kurie domisi kalbos technologijomis.

## ANOTUOTI LIETUVIŲ KALBOS TEKSTYNAI IR JŲ SUDARYMO YPATUMAI

### 1. Morfologiškai anotuotas tekstynas MATAS

Morfologinė analizė yra pirmasis kalbos apdorojimo etapas ir neretai paruošia tekstą tolesnei sintaksinei ar semantinei analizei, todėl labai svarbu, kad morfologiniai anotatoriai veiktų kuo tiksliau, nes morfologijos lygmenyje padarytos klaidos trukdo automatinei kitų lygmenų kalbos analizei.

Pirmasis morfologinis lietuvių kalbos anotatorius sukurtas maždaug prieš 25 metus. Jo autorius – Vytautas Zinkevičius (plačiau žr. Zinkevičius 2000). Šiuo metu lietuvių kalbai viešai prieinami du morfologiniai anotatoriai: minėta V. Zinkevičiaus programa, dažniausiai vadinama *Lemuokliu*[5] (nuo žodžio *lema* – antraštinė forma), ir portale *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema* (toliau semantika.lt) prieinamas anotatorius[6]. Šiame straipsnyje pristatomas tekstynas MATAS anotuotas naudojant *Lemuoklį*, o sintaksiškai anotuotas tekstynas rengtas naudojant semantika.lt morfologinį anotatorių.

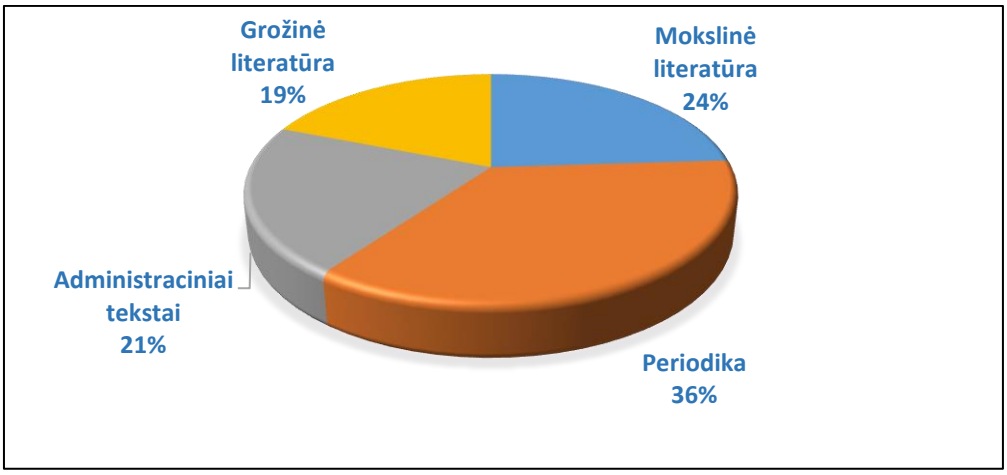
Nors *Lemuoklis* veikia gana gerai, bet šios programos negalima atnaujinti, nes ji yra uždaro kodo. Semantika.lt anotatorius sukurtas *Hunspell* platforma kaip atvirojo kodo programa; šis anotatorius yra atnaujinamas, papildomas naujais žodžiais (dabar apima 171 000 lemų), tvarkomos jo taisyklės (plačiau Dadurkevičius 2017). 2017 m. atlikus tyrimą ir įvertinus abiejų anotatorių kokybę, nustatyta, kad tiksliau veikia semantika.lt anotatorius (Kapočiūtė-Dzikienė ir kt. 2017).

Toliau pristatomas tekstynas MATAS rengtas 2002–2014 m. Kaip pagrindas imtas 1 mln. žodžių tekstynas, sudarytas 2006 m., jis papildytas naujais tekstais, pertvarkytas. Pirmiausia MATO tekstai buvo anotuoti automatiškai morfologiškai, vėliau jie kalbininko peržiūrėti ir sutvarkyti.

[5] Prieiga internete: http://tekstynas.vdu.lt/page.xhtml?id=morphological-annotator; ši programa ne tik morfologiškai anotuoja tekstą, bet gali ir sintezuoti, t. y. sugeneruoti reikalingas formas; ji pritaikyta ir seniesiems raštams analizuoti.

[6] Prieiga internete: http://semantika.lt/SyntaticAndSemanticAnalysis/Analysis.

MATĄ sudaro 1,6 mln. tekstų, sudarytų iš dokumentų, grožinės literatūros kūrinių, mokslinių tekstų. Kaip ir *Dabartinės lietuvių kalbos tekstyne*, didžiausią dalį (36 proc.) sudaro publicistikos tekstai (žr. 1 paveikslą).




1 PAV. Morfologiškai anotuoto tekstyno MATAS sandara (procentais)

MATAS prieinamas keliais formatais: vadinamuoju KLC formatu, kur naudojami kiek neįprasti kalbos dalių ir gramatinių kategorijų sutrumpinimai (šiuo formatu anotuotas ir http://clarin-lt.lt/ portalo saugykloje prieinamas tekstynas), pvz.:

```
<word="Tarp" lemma="tarp" type="prln">
<space>
<word="tankiai" lemma="tankiai" type="prvks teig nelygin.l">
<space>
<word="suaugusių" lemma="suaugti(-ga,-go)" type="dlv teig nesngr veik.r būt.kart.l neįvardž vyr.gim dgsk K">
<space>
<word="medžių" lemma="medis" type="dktv vyr.gim dgsk K">
<space>
<word="kur ne kur" lemma="kur ne kur" type="prvks teig nelygin.l">
<space>
<word="vis" lemma="vis" type="prvks teig nelygin.l">
<space>
<word="pasimatydavo" lemma="pasimatyti(-to,-tė)" type="vksm teig sngr tiesiog.nuos būt.d.l vnsk IIIasm">
<space>
<word="dangaus" lemma="dangus" type="dktv vyr.gim vnsk K">
<space>
<word="lopinėlis" lemma="lopinėlis" type="dktv vyr.gim vnsk V">
<sep=".">
<p>
```

MATAS taip pat anotuotas tarptautiniu TEI P5 formatu. Juo kalbos dalys ir gramatinės kategorijos trumpinamos viena raide arba skaitmeniu, pvz., *vatstdv3* reiškia atitinkamai:

veiksmažodis, asmenuojamoji forma, teigiamoji forma, sangrąžinis, tiesioginė nuosaka, būtasis dažninis laikas, vienaskaita, trečiasis asmuo. Toliau pateikta ta pati teksto ištrauka TEI P5 formatu:

```
<w lemma="tarp" ana="#r">Tarp</w>

<pc> </pc>
<w lemma="tankiai" ana="#ptn">tankiai</w>
<pc> </pc>
<w lemma="suaugti(-ga,-go)" ana="#vdtnvknvdk">suaugusių</w>
<pc> </pc>
<w lemma="medis" ana="#dbvdk">medžių</w>
<pc> </pc>
<w lemma="kur ne kur" ana="#ptn">kur ne kur</w>
<pc> </pc>
<w lemma="vis" ana="#ptn">vis</w>
<pc> </pc>
<w lemma="pasimatyti(-to,-tė)" ana="#vatstdv3">pasimatydavo</w>
<pc> </pc>
<w lemma="dangus" ana="#dbvvk">dangaus</w>
<pc> </pc>
<w lemma="lopinėlis" ana="#dbvvv">lopinėlis</w>
<pc>.</pc>
```

Dar vienas formatas, kuriuo morfologiškai anotuoti tekstai prieinami http://corpus.vdu.lt svetainėje, sudarytas remiantis *MULTEXT-East*[7] formato pavyzdžiu. Pagal šį formatą kiekviena kalbos dalis turi skirtingą morfologinių kategorijų skaičių (nuo 2 iki 14), jos užrašomos santrumpomis, pvz., pažymoje *Ncmpnn- N* reiškia daiktavardį, *c* – bendrinį daiktavardį, *m* – vyriškąją giminę, *p* – daugiskaitą, *n* – vardininką, *n* – nesangrąžinį daiktavardį, gale brūkšnelis žymi, kad šiam žodžiui, pvz., *universitetai*, nepriskiriama jokia semantinė pažyma[8].

Trečiame skyriuje pristatysime paiešką tiek MATO, tiek ALKSNIO tekstynuose naudojant ANNIS sistemą. Joje naudojama dar vienas morfologinių pažymų formatas, jis sudarytas pagal Leipcigo glosavimo pažymas (plačiau žr. 3 skyrių).

Skirtingos anotavimo sistemos būdingos ne tik lietuvių kalbos technologijoms. Ši problema paskatino sukurti konvertavimo įrankį *Pepper* (Zipser ir kt. 2010). Tai rodo įvairių išteklių kūrimo etapus ir naudojamus anotavimo įrankius.

Kaip matyti iš anksčiau pateiktų pavyzdžių, KLC naudojamos keturios morfologinių pažymų sistemos. Keli skirtingi anotavimo formatai atsirado dėl to, kad morfologiškai anotuotas tekstynas rengtas ilgą laiką. Vienu metu iš viso nebuvo taikomi jokie anotavimo standartai, tiesiog buvo naudojami morfologinio analizatoriaus kūrėjo V. Zinkevičiaus sukurti kalbos dalių ir gramatinių kategorijų sutrumpinimai. Jie lengvai suprantami vartotojams. Po kurio laiko dėl tarptautinio bendradarbiavimo morfologiškai anotuotas tekstynas peranotuotas naudojantis

[7] Prieiga internete: http://nl.ijs.si/ME/V4/msd/html/index.html.
[8] Plačiau apie šias pažymas žr. http://corpus.vdu.lt/lt/morph.

tarptautiniais standartais. Kaip minėta, semantika.lt (šios sistemos pagrindinis tikslas buvo sukurti internetines paslaugas) morfologiniame analizatoriuje naudojamos pažymos *MULTEXT-East* formatu. Pažymos trumpos, bet žmonėms sunkiai suprantamos, jos tinkamesnės kompiuterinei analizei.

*MULTEXT-East* formatu morfologiškai anotuotas ir sintaksinis tekstynas ALKSNIS. Tekstynų paieškos įrankis ANNIS (žr. 3 skyrių) suteikė galimybę sujungti ALKSNIO ir MATO morfologinio anotavimo formatus. Pirma mintis buvo naudoti *Universal Dependency* (UD) pažymas, nes tai gana paplitusi anotavimo sistema ir ji lengvai skaitoma (anglų kalba). Tačiau UD pažymos labai ilgos ir tai pasirodė nepatogu atliekant paiešką per ANNIS: kiekvieno žodžio pažymos užima daug vietos, todėl, analizuojant ieškomo žodžio kontekstą, reikėjo slinkti lango rodyklę į vieną ar kitą pusę. Dėl šios praktinės problemos nuspręsta adaptuoti Leipcigo glosavimo pažymas, nes jos lengvai skaitomos ir gana trumpos. Nepaisant čia aprašytų anotavimo pažymų įvairovės, pabrėžtina, kad abiejų tekstynų ANNIS sąsajoje naudojama tik viena (adaptuota Leipcigo glosavimo) sistema, o tai turėtų būti didelis palengvinimas vartotojams.

Paminėtina, kad MATE nenurodytos sakinių ribos, tą galima padaryti greitai automatiškai, bet reikia patikrinti gautus rezultatus. MATAS, kaip ir toliau aptarsimas tekstynas ALKSNIS, prieinamas http://clarin-lt.lt/ portalo saugykloje[9]. Jame galima atlikti paiešką naudojant ANNIS sistemą[10] (plačiau žr. 3 skyrių).

## 2. Sintaksiškai anotuotas tekstynas ALKSNIS

Lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS parengtas 2016 m. Jo rengimą rėmė Bendrosios kalbos išteklių ir technologijų infrastruktūra Europos mokslinių tyrimų infrastruktūros konsorciumas (CLARIN ERIC) (MTI-02/2015). Tekstynas prieinamas viešai http://clarin-lt.lt/ portalo saugykloje[11]. Kaip juo naudotis, paaiškinta vartotojo gide[12].

Tekstynas suplanuotas kaip lietuviškosios sintaksinės analizės etalonas. Tokio pobūdžio tekstynams būdinga tai, kad jie nedideli, optimalios žanrinės sudėties, o sintaksinė analizė, nors ir atlikta automatiškai, bet peržiūrėta ir ištaisyta kalbininkų. Remiantis sintaksinės analizės etalonu, galima kurti statistiniu pagrindu veikiantį sintaksinį analizatorių (angl. *statistically-based parser*) ir plėsti automatinę sintaksinę analizę (Bielinskienė ir kt. 2016).

[9] Prieiga internete: https://clarin.vdu.lt/xmlui/handle/20.500.11821/9.
[10] Prieiga internete: http://158.129.51.247:8080/annis-gui-3.4.4.
[11] Prieiga internete: https://clarin.vdu.lt/xmlui/handle/20.500.11821/10.
[12] Prieiga internete: https://youtu.be/PlE0PWurb4Y.

Kadangi ALKSNIS yra naujesnis lietuvių kalbos išteklius nei MATAS, mažiau aprašytas, todėl šiame straipsnyje ALKSNIUI skirsime didesnį dėmesį, taip pat aptarsime sintaksiškai anotuojant kilusius klausimus.

## 2.1. Tekstyno dalys

Tekstyną ALKSNIS sudaro 30 599 žodžiai, 2355 sintaksiškai anotuoti sakiniai. Lyginant su artimų šalių kalbų tekstynais, šis tekstynas gana panašus ir, galima teigti, yra pakankamo dydžio: latvių kalbos tekstyną[13] sudaro 3800 sakinių, 53 tūkst. žodžių (Predkalnina ir kt. 2016); estų kalbos [14] – 1400 sakinių, 10 600 žodžių (Muischnek ir kt. 2014); lenkų kalbos [15] – 8227 sakinių. Pirmąjį sintaksiškai anotuotą tekstyną PENN (JAV) sudaro gerokai daugiau – apie 3 mln. žodžių[16].

ALKSNIO tekstai apima keturias žanrines dalis (žr. 2 pav.). Bendrosios bei specialiosios periodikos ir grožinės literatūros dalys yra apylygės, o administracinių tekstų dalis mažiausia, nes tai specifinė lietuvių kalba: sakiniai dažnai būna ilgi, juose gausu dokumentų pavadinimų, nuorodų į kitus dokumentus ir pan. Tokių sakinių anotavimas neatskleistų tipiškos sintaksinės struktūros.


2 PAV. Sintaksiškai anotuoto tekstyno ALKSNIS sandara

Tekstynui sudaryti imti ištisi, nesutrumpinti tekstai. Sakiniai atrinkti iš kuo įvairesnių tekstų žanrų. Grožinės literatūros dalį sudaro sakiniai iš lietuvių autorių kūrinių. Atrinkti sakiniai parengti kompiuterinei analizei: patikrintas teksto kodavimas, ištrintos lentelės ir paveikslai.

[13] Prieiga internete: http://sintakse.korpuss.lv/index.html.
[14] Prieiga internete: https://metashare.ut.ee/repository/browse/estonian-treebank/4eb86e5e463411e2a6e4005056b400242c46832883754ad7bd89d07f69d6a0fc/.
[15] Prieiga internete: http://zil.ipipan.waw.pl/Sk%C5%82adnica.
[16] Prieiga internete: https://catalog.ldc.upenn.edu/ldc99t42.

## 2.2. Tekstyno anotavimo ir vizualizavimo įrankiai

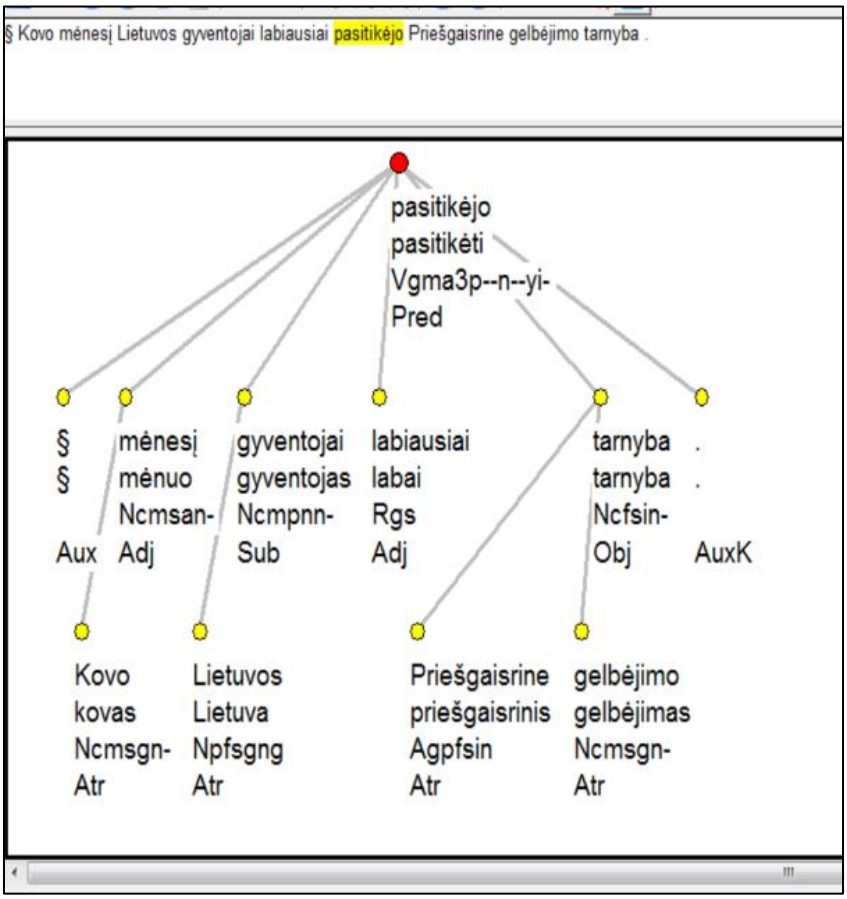
Automatinė sintaksinė analizė remiasi VDU KLC projekto „Lietuvių kalbos sintaksinės-semantinės analizės sistema tekstynui, lietuviškam internetui ir viešojo sektoriaus taikymams“ metu parengtu sintaksiniu analizatoriumi, kuris sukurtas *Haskell* kalba[17]. Analizatorius remiasi taisyklėmis pagrįstu (angl. *rule-based*) metodu. Šis įrankis skaito semantika.lt segmentavimo ir morfologinės analizės modulių morfologiškai anotuotus failus ir kaip rezultatą pateikia sakinius, suskaidytus į dėmenis ar sintagmas, generuoja sintaksinius medžius, nurodo jų teminius vaidmenis (angl. *thematic roles*) ir sintaksines funkcijas (Boizou ir kt. 2014: 69).

Internetinės paslaugos generuoja tiktai JSON failus, bet kitas KLC įrankis konvertuoja sintaksinės analizės rezultatus PML (*Prague Markup Language*) formatu. Sintaksiškai anotuotam tekstynui yra generuojami priklausomybių medžiai (angl. *dependency trees*). Šis formatas leidžia vizualizuoti ir redaguoti PML formatu priklausomybių medžius naudojant Prahos Karolio universiteto UFAL sukurtą TrED redaktorių[18] (žr. 3 pav.). Visi automatiškai suanotuoti sakiniai patikrinti ir ištaisyti kalbininkų. Taigi sintaksinė analizė apima tris lygmenis: automatinį žodžių ir sakinių segmentavimą bei morfologinę analizę, automatinę sintaksinę analizę (taisyklėmis pagrįstu metodu) ir rankinį sakinių tvarkymą (plačiau žr. 2.4).

## 2.3. ALKSNIO duomenų struktūra

Atlikus automatinę sintaksinę analizę, sakiniai pateikiami grafiškai medžio principu (žr. 3 pav.). Kiekviena medžio viršūnė atitinka sakinio žodį, skyrybos ženklą ar kitą sakinio vienetą (simbolį, skaitmenį ir pan.). Priklausomybių ryšiai tarp žodžių yra nurodomi „šakomis“ arba briaunomis.

[17] Sintaksinis analizatorius sukurtas VDU Kompiuterinės lingvistikos centre L. Boizou ir F. Zamblera'os.
[18] Prieiga internete: https://ufal.mff.cuni.cz/tred/.




3 PAV. Sintaksiškai anotuoto sakinio priklausomybių medis TrEd redaktoriaus formatu

Prie visų žodžių tokia eilės tvarka nurodoma (žr. 3 pav.):

1) konkreti sakinyje pavartota žodžio forma (pvz., *gyventojai*);
2) antraštinė (žodyninė forma) – lema (pvz., *gyventojas*);
3) morfologinės pažymos (pvz., *Ncmpnn-*);
4) sintaksinė funkcija (pvz., *Sub*) (sutrumpinimų paaiškinimus žr. toliau).

## 2.4. Analizės lygmenys

*Morfologinės analizės lygmuo.* Morfologinis anotavimas yra pirmasis automatinės analizės etapas, todėl yra būtinas aukštesnio lygmens sintaksinei arba sakinio analizei. Morfologinė analizė atlikta naudojant semantika.lt anotatorių (Dadurkevičius 2017). ALKSNYJE naudojamos morfologinės pažymos sudarytos remiantis *MULTEXT-East* formato pavyzdžiu (plačiau aprašyta 1 skyriuje).

*Sintaksinės analizės lygmuo.* Sintaksinis analizatorius apdoroja morfologiškai anotuotus failus ir pateikia sakinius, suskaidytus į dėmenis arba sintagmas, sugeneruoja sintaksinius medžius, nurodo jų elementų sintaksines funkcijas. Analizė remiasi priklausomybių gramatikos modeliu (angl. *dependency grammar*). Jis paprastai taikomas tipologiškai panašioms kalboms, kurioms būdinga žodžių formų kaityba ir laisva žodžių tvarka, pvz., slavų kalboms (PDT[19], SynTagRus[20]), latvių kalbai ir t. t. Sintaksinės priklausomybės išreiškiamos pagal hierarchiją. Pradedama nuo pagrindinio sakinio dėmens – medžio viršūnės, prie kurios jungiami kiti sakinio dėmenys pagal jų priklausomumą (t. y. sintaksinius ryšius).

Sintaksinių pažymų sutrumpinimai ir informacija apie sintaksinius ryšius bei priklausomybes paremti čekų įdirbiu (Hajič ir kt. 1999); jie vieni pirmųjų išplėtojo sintaksinę analizę sintetinio tipo čekų kalbai. Šiuo metu ALKSNYJE yra naudojama 18 pagrindinių sintaksinių pažymų (neskaičiuojant jų variantų): *Sub* – subjektas, *Pred* – tarinys (predikatas), *Obj* – objektas ir t. t. (žr. 1 lentelę). Pažymų variantų atsiranda tada, kai į vieną sudedamos dvejopos arba dvigubos pažymos, pvz., žymint sudėtinį vardažodinio arba veiksmažodinio tipo tarinį naudojamos dvejopos pažymos: atitinkamai *PredN* ir *PredV*; žymint sudėtinio sakinio šalutinį pažyminio (atributinį) dėmenį, prie šalutinio sakinio dėmens tarinio žymima dviguba pažyma *Pred_Atr*; norint parodyti sujungimo ryšį (koordinaciją) prie vienarūšių sakinio dėmenų kaip dvigubos pažymos dalis žymima *_Co*, t. y. prie kiekvieno iš išvardytų vienarūšių papildinių rašoma *Obj_Co* ir pan. Beje, sudėtiniuose sakiniuose gali atsirasti ir trigubų pažymų, pavyzdžiui, kai reikia pažymėti vienarūšius šalutinius sakinius: *Pred_Atr_Co* – tai šalutinis pažyminio (atributinis) sakinys.

1 LENTELĖ. Sintaksinės funkcijos ir jų pažymos tekstyne

| Pažyma | Sintaksinė funkcija | Pavyzdys |
|---|---|---|
| **Sub** | Subjektas | *Jis [Sub] sakė* |
| **Pred** | Predikatas (arba pagalbinis žodis) | *Jis ėjo [Pred]; buvo [Pred] patenkintas* |
| **PredN** | Vardažodinė predikato dalis | *Buvo [Pred] patenkintas [PredN]* |
| **PredV** | Veiksmažodinė predikato dalis | *Turi [Pred] atsilyginti [PredV]* |
| **Obj** | Objektas | *Laukiu svečio [Obj]; reikia spręsti [Obj]* |
| **Atr** | Atributas (pažyminys) | *Vidurinė [Atr] mokykla; šalis narė [Atr]* |
| **Adj** | Aplinkybės | *Dabar [Adj] nuspręs; Partizanų gatvėje [Adj]* |

[19] Prahos priklausomybių medžių bankas (*Prague Dependency Treebank*), plačiau žr. https://ufal.mff.cuni.cz/pdt3.0.

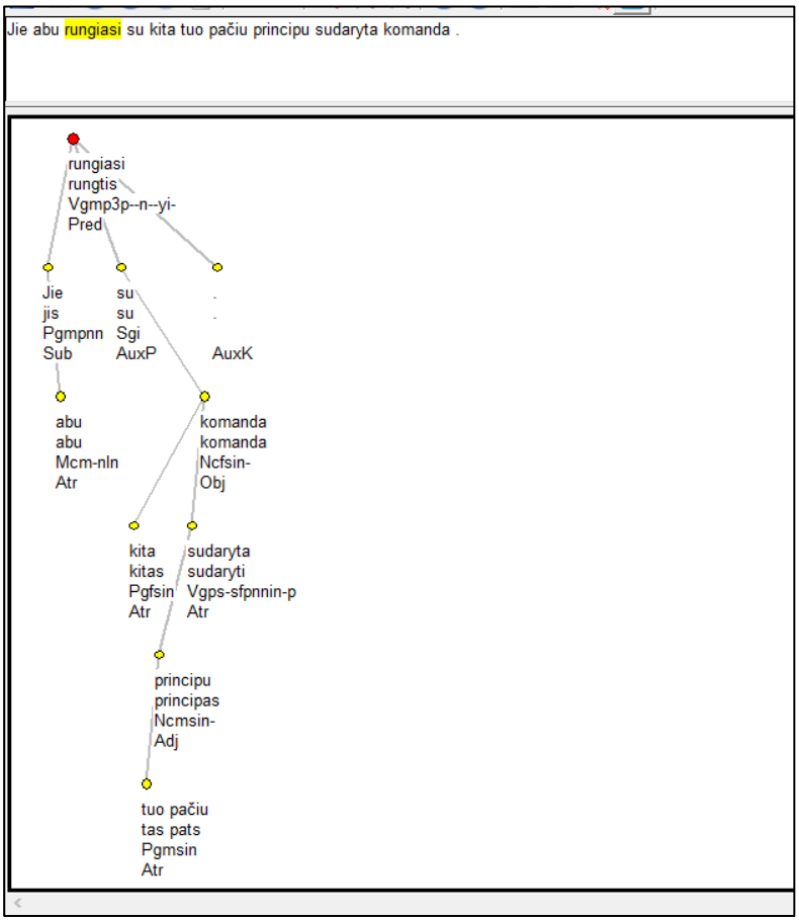
[20] Prieiga internete: http://www.ruscorpora.ru/en/.

| | | |
|---|---|---|
| **Aux** | Pagalbinė funkcija (pvz., kableliai ar kiti simboliai) | |
| **AuxC** | Subordinacija (prijungiamieji sakiniai) | *Sakė, kad [AuxC]...* |
| **AuxK** | Sakinio pabaigos skyrybos ženklas | |
| **AuxL** | Pagalbinis leksinis vienetas (pvz., žodis užsienio kalba) | *JAV kompanija* North [AuxL] American [AuxL] Investment [AuxL] Consulting [AuxL] Inc. [Atr] |
| **AuxP** | Prielinksnis | *Pereis į [AuxP] lygį* |
| **AuxZ** | Dalelytė | *Kaip ir [AuxZ] visada* |
| **Coord** | Sujungimas | *Gamintojų ir [Coord] pardavėjų* |
| **_Co (pvz., Atr_Co)** | Sujungiamieji dėmenys (pvz., vienarūšės sakinio dalys) | *Gamintojų [Atr_Co] ir [Coord] pardavėjų [Atr_Co]* |
| **Par** | Įterpiniai, kreipiniai | *Tiesą sakant [Par], atrodo* |
| **ExD** | Elipsė ar kiti praleidimai sakinyje | *Moteris – [Pred_ExD] būtybė* |
| **Pred_ (pvz., Pred_Obj)** | Prijungiamojo sakinio šalutinio dėmens tipas (žymima prie predikato) | *Rašė, kad ypatybes lemia [Pred_Obj]* |

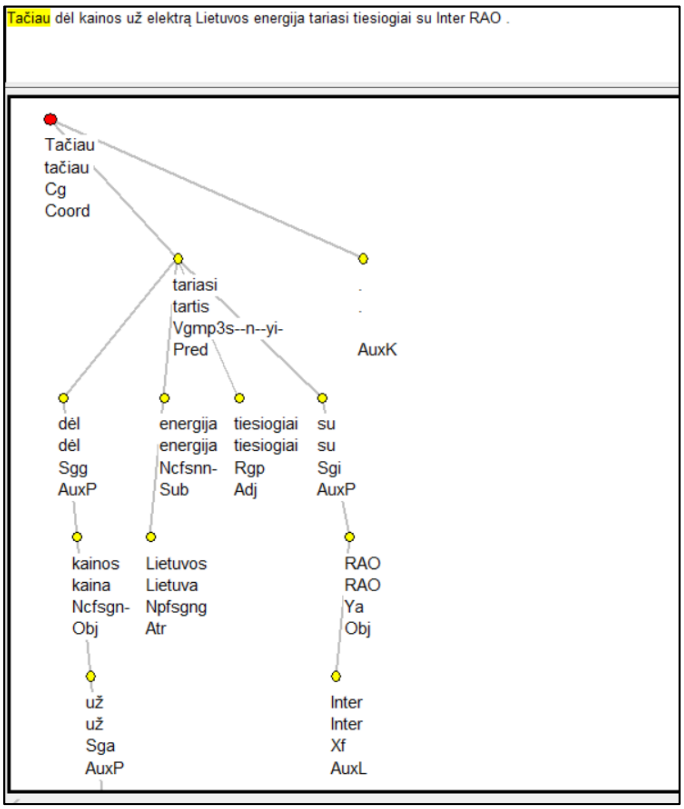
*Rankinio sakinių tvarkymo lygmuo.* Visi automatiškai sintaksiškai anotuoti sakiniai buvo patikrinti kalbininkų. Daugelio sakinių, ypač sudėtinių, anotavimą reikėjo taisyti ir nurodyti taisyklingas priklausomybes. Kai kur taisytos ir morfologinės pažymos. Sakinius atskirai taisė trys kalbininkai, o vėliau kiekvieną sakinį dar papildomai patikrino visi trys kalbininkai. Sintaksinio anotavimo ypatumai aprašomi toliau.

*Sintaksinio anotavimo ypatumai.* Atliekant automatinę sintaksinę analizę, kaip minėta, sintaksinio medžio viršūne (angl. *root*) laikomas hierarchiškai aukščiausias sakinio dėmuo (pvz., veiksmažodis, kartais sakinio pradžios jungtukas) ar kitas sakinio elementas (pvz., kai yra sudėtinis sakinys – jo jungukas ar jungiamasis žodis) ir kaip atšaka, rodanti sakinio skyrybą, – sakinio skyrybos ženklas: taškas, klaustukas, šauktukas, daugtaškis, kabliataškis ar kt. Toliau pateiksime konkrečių pavyzdžių, kad iliustruotume sintaksinio anotavimo ypatumus.

Pavyzdys, kai sakinio viršūnė – veiksmažodis [*Pred*], parodytas 4 paveiksle, o kai jungtukas [*Coord*] – 5 paveiksle.

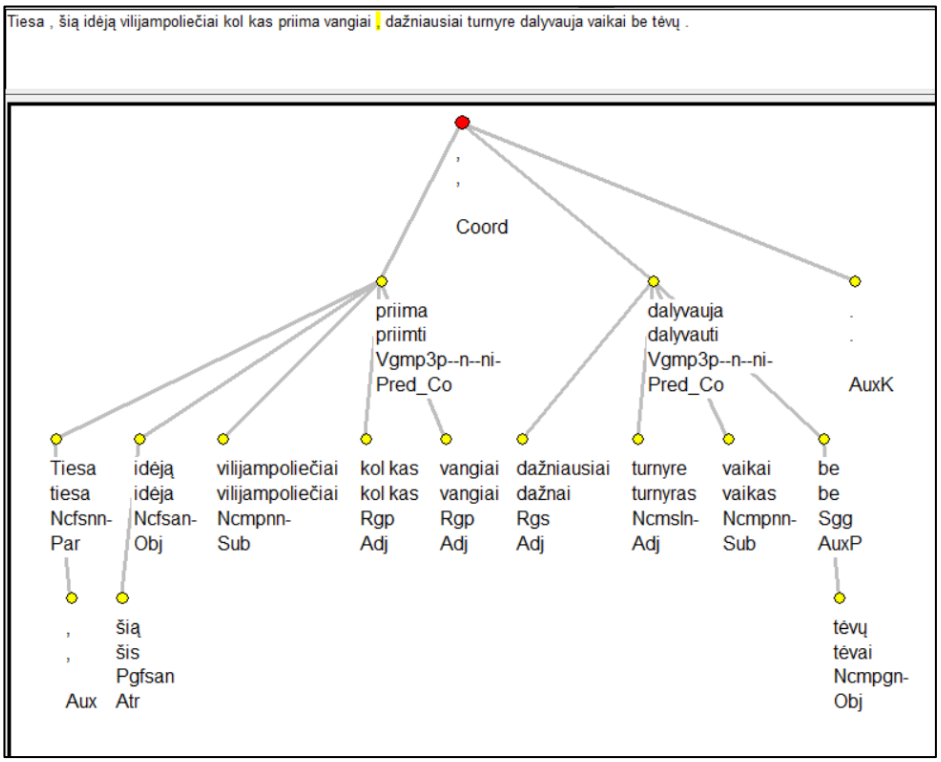



4 PAV. Priklausomybių medžio viršūnė – veiksmažodis



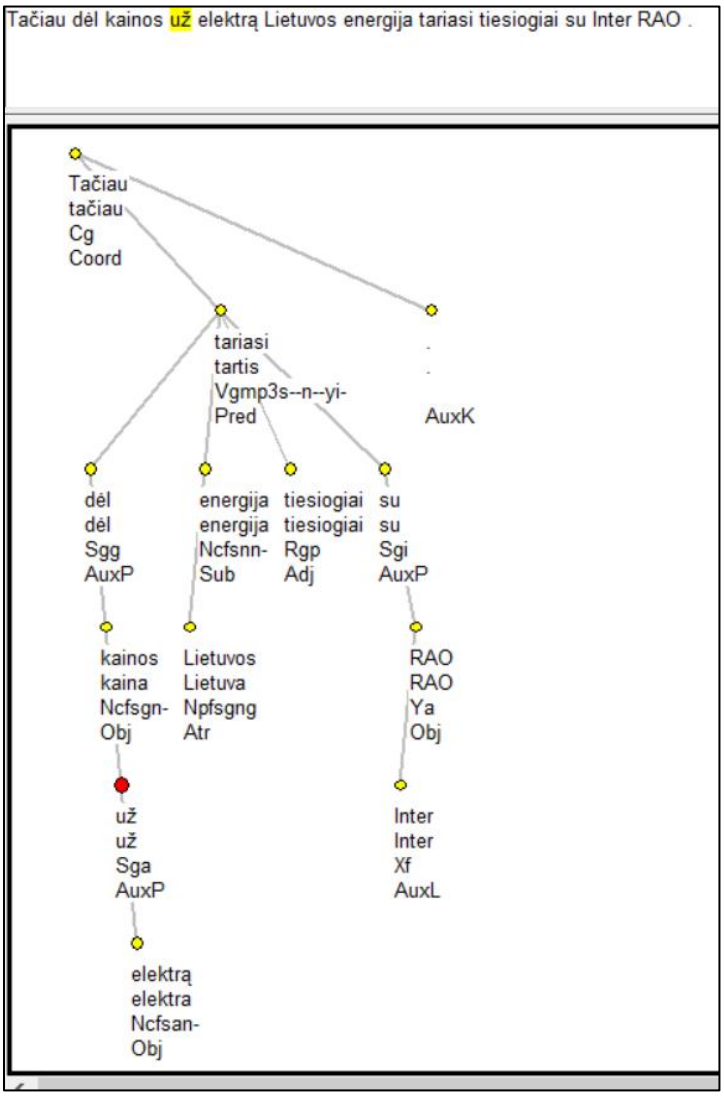

5 PAV. Priklausomybių medžio viršūnė – sakinio pradžios jungtukas

Pavyzdys, kai sudėtinio sakinio viršūnė – skyrybos ženklas, pateiktas 6 paveiksle.



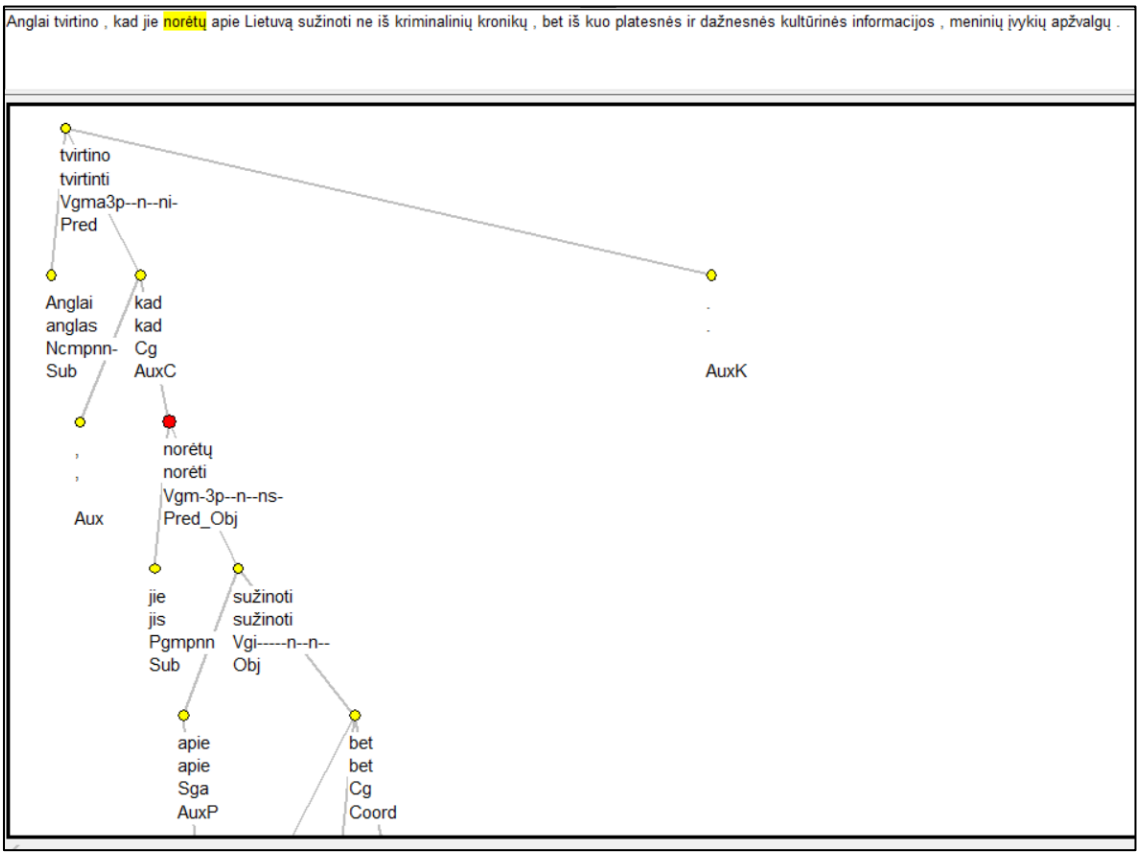

6 PAV. Priklausomybių medžio viršūnė – sudėtinio sakinio dėmenų skyrybos ženklas

Laikoma, kad prielinksninėse konstrukcijose prielinksnis užima aukštesnę poziciją, jis žymimas *AuxP*, o po jo einantis žodis žymimas pagal atliekamą funkciją (pvz., objektas): *kainos už* [*AuxP*] *elektrą* [*Obj*] (žr. 7 pav.).



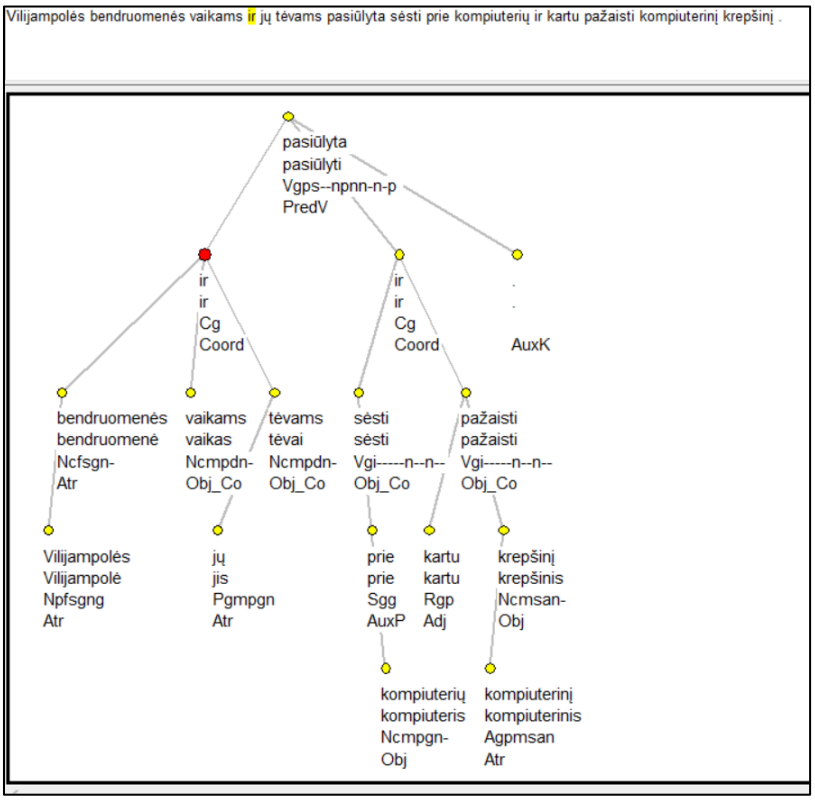

7 PAV. Prielinksninių konstrukcijų anotavimas

Šalutiniai sudėtinio sakinio dėmenys priklauso nuo pagrindinio sakinio dėmens tarinio, o nuo jo priklauso šalutinio sakinio dėmens jungtukas (kablelis priklauso nuo jungtuko). Svarbu nurodyti ir šalutinio sakinio dėmens funkciją. Ji nurodoma prie šalutinio sakinio dėmens tarinio, pvz., *Pred_Obj* – šalutinis papildinio (objektinis) sakinys (žr. 8 pav.).



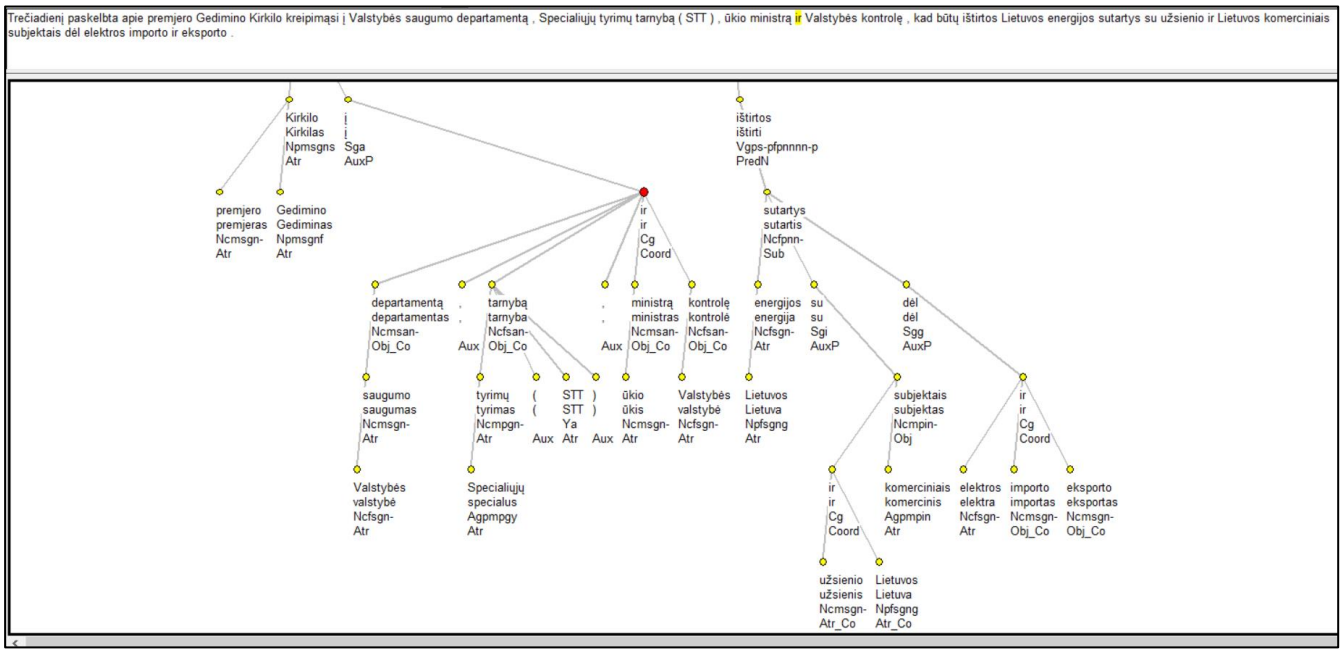

8 PAV. Šalutinių prijungiamojo sakinio dėmenų anotavimo pavyzdys

Sujungiamojo ryšio rodiklis (jungtukas arba kablelis) gali būti sujungiamosios konstrukcijos viršūnė (pvz., *pasiūlyta sėsti ir* [Coord] *pažaisti*; *pasiūlyta vaikams ir* [Coord] *jų tėvams*) (žr. 9 pav.).



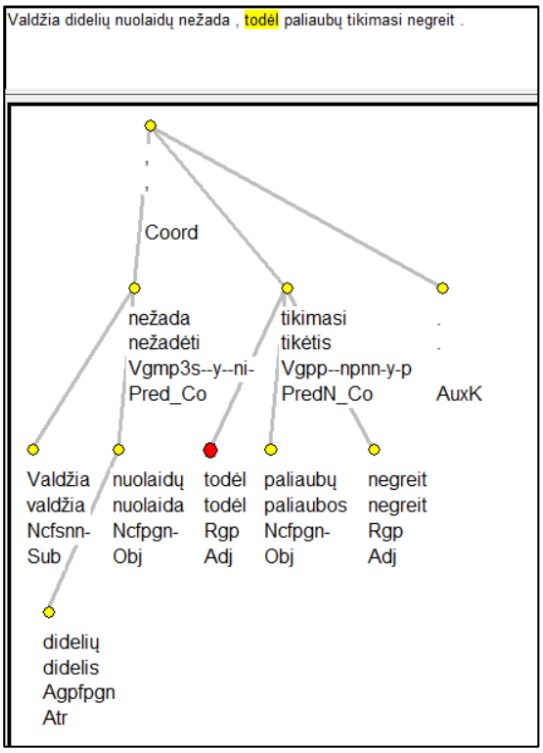

9 PAV. Sujungiamuoju ryšiu susijusių dėmenų anotavimo pavyzdys

Jei yra daugiadėmenis mišrusis sujungimas, tai konstrukcijoje esantis jungtukas yra viršūnė, o nuo jo eina visi kiti bejungtukiai dėmenys (kableliai su savo išvardijamaisiais žodžiais yra tame pačiame priklausomybės lygmenyje). Kiti skyrybos ženklai žymimi *Aux* (žr. 10 pav.).

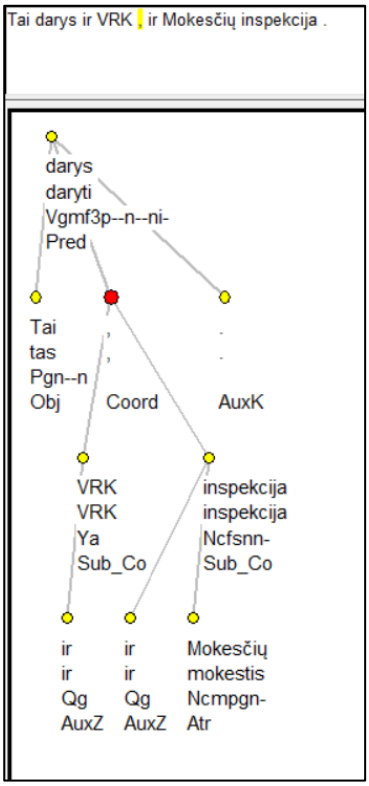



10 PAV. Mišriosios sujungiamosios konstrukcijos anotavimo pavyzdys (paveiksle matyti ir daugiau sujungiamojo ryšio atvejų)

Sujungiamųjų sakinių viršūnė yra sujungiamasis jungtukas. Jungiamieji žodžiai, kuriuos sudaro prieveiksmiai, įvardžiai ar dalelytės (pvz., prieveiksmis *todėl*), priklauso nuo antrojo sakinio dėmens predikato ir atlieka aplinkybės funkciją (*Adj*), o sujungimo *Coord* funkciją atlieka kablelis kaip ir bejungtukiuose sakiniuose arba sakiniuose su kartojamaisiais jungtukais, kurie čia laikomi dalelytėmis (žr. 12 pav.), kaip ir sakinio dėmenų sujungimo atveju (žr. 11, 12 pav.).

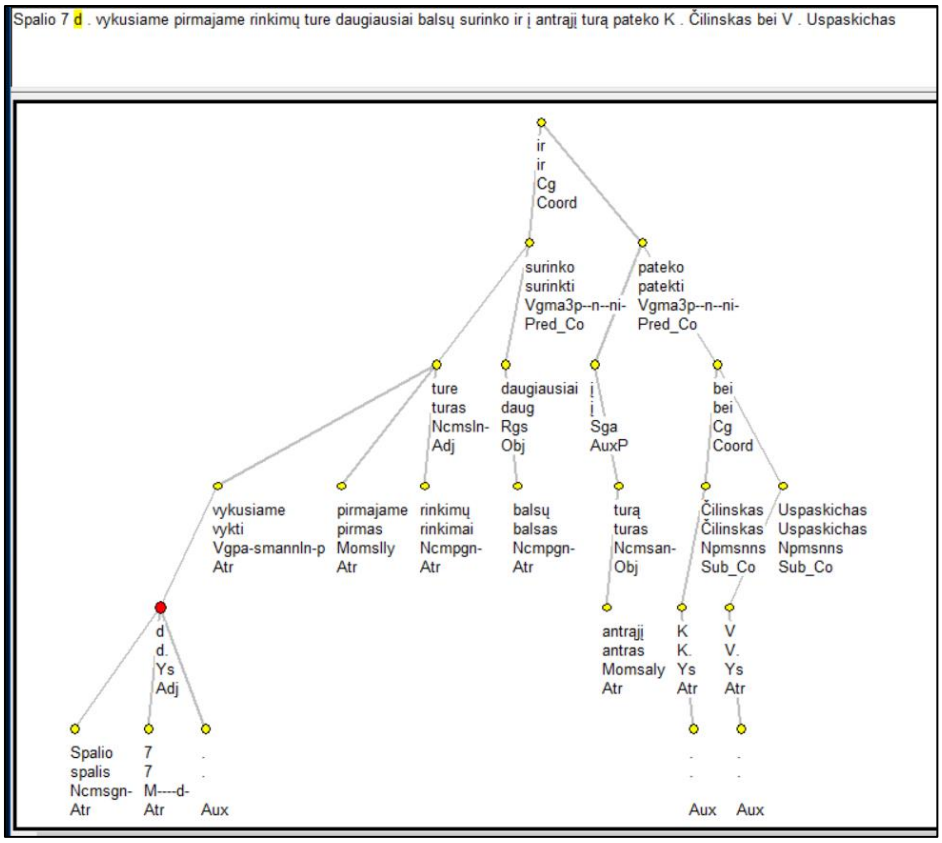



11 PAV. Sujungiamojo sakinio su jungiamuoju žodžiu anotavimo pavyzdys

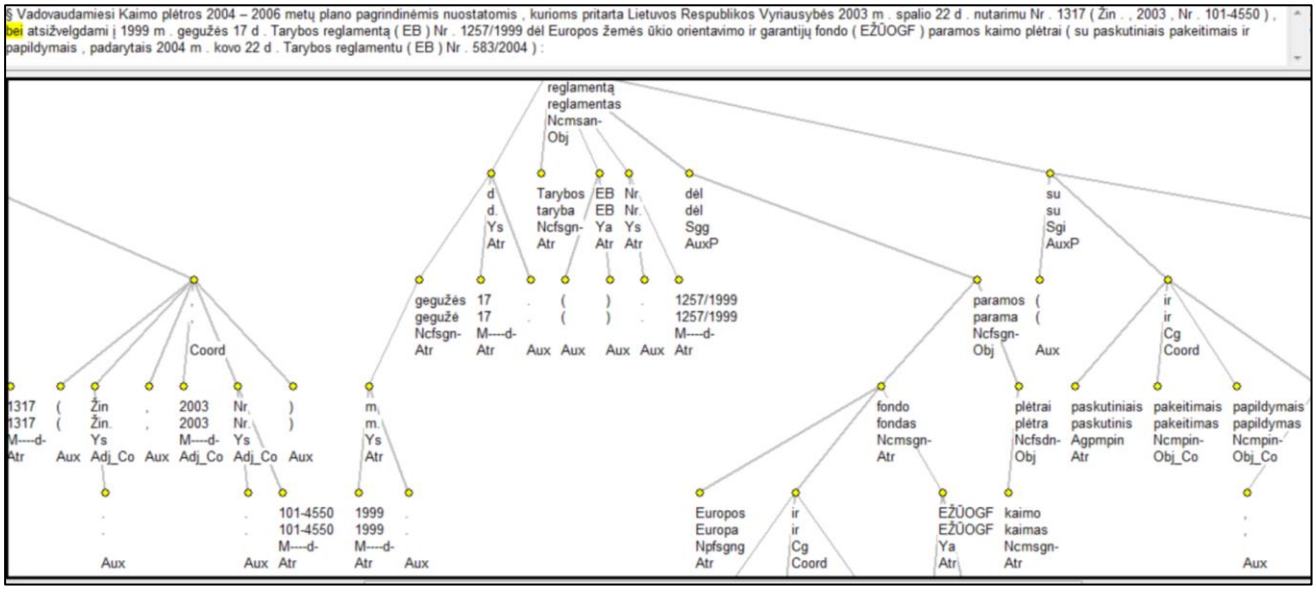



12 PAV. Sujungiamosios konstrukcijos anotavimas

Sudėtinga anotuoti skaitines išraiškas, pvz., valandas, metus. Jos anotuotos kaip aplinkybės su pažyminiais arba kaip pažyminiai (žr. 13, 14 pav.).



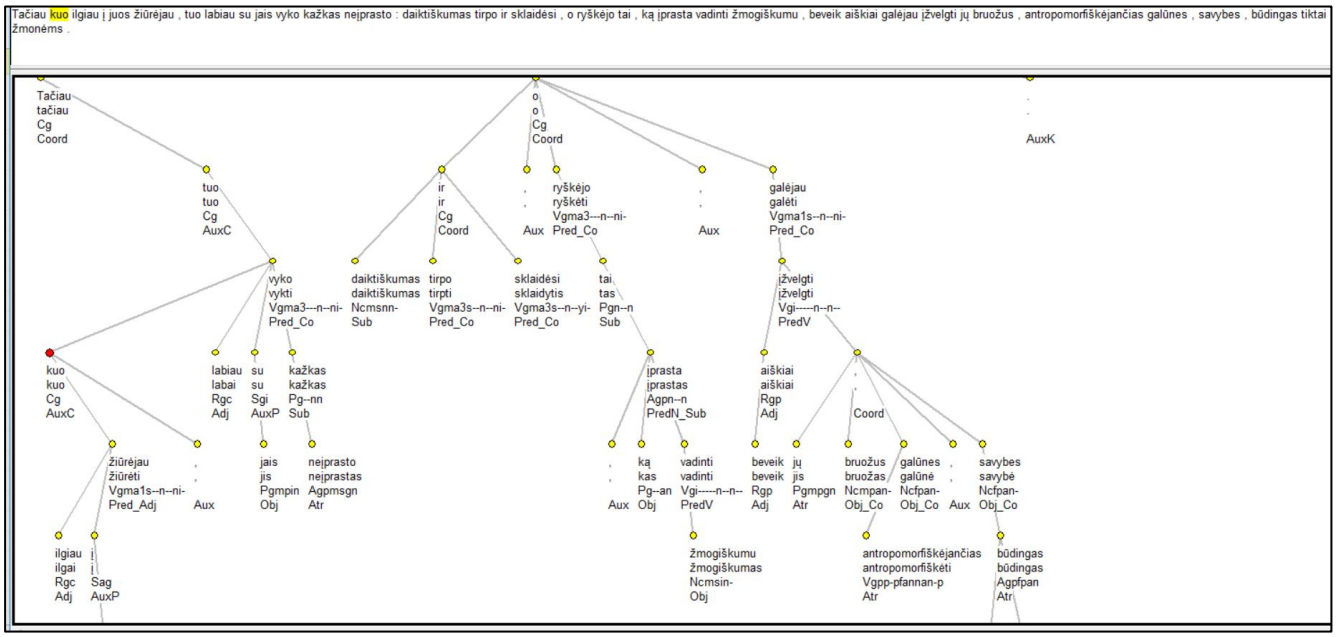

13 PAV. Datos anotavimas




14 PAV. Datų ir kitų skaitinių išraiškų anotavimas (paveiksle pateiktas tik sintaksiškai anotuoto sakinio fragmentas)

Ne visada lengva anotuoti sudėtinius tarinius. Vienais atvejais jie anotuojami taip, kaip ir suprantami lietuvių kalbos gramatikose, t. y. kaip vardažodinio ir veiksmažodinio tipo tariniai (pvz., *yra [Pred] paprastas [PredN]; gali [Pred] būti [PredV]*), o kai kuriais atvejais antrasis tarinio dėmuo laikomas objektu (pvz., *ketinu [Pred] sukviesti [Obj]*).

Buvo sudėtinga anotuoti konstrukcijas su poriniais jungtukais, nes abu jungtukai hierarchiškai yra lygiaverčiai. Tokiu atveju laikyta, kad aukštesnę hierarchijos poziciją užima pagrindinio sakinio dėmens tarinys, nuo jo priklauso šalutinio sakinio dėmuo (žr. 15 pav.).




15 PAV. Sudėtinių sakinių su poriniais jungtukais anotavimo pavyzdys

Galiausiai svarstyta dėl kai kurių kalbos dalių statuso. Pavyzdžiui, kartojamuosius sujungiamojo ryšio jungtukus nuspręsta laikyti dalelytėmis, nes jų funkcija priklausomybių medžio struktūroje panaši į dalelyčių (žr. 12 pav.).

Anotuojant sintaksiškai rankomis buvo sprendžiama nemažai problemų, čia jų buvo įvardyta tik keletas. Daugelis anotavimo sunkumų išspręsta laikantis tradicinės gramatikos požiūrio, tik susitariant dėl tam tikro anotavimo būdo. Kitais atvejais teko prisitaikyti prie sintaksinio anotavimo ypatumų ir anotuojant nuosekliai laikytis susitarimo. Tikimasi iškilusias problemas spręsti toliau plėtojant automatinę sintaksinę analizę.

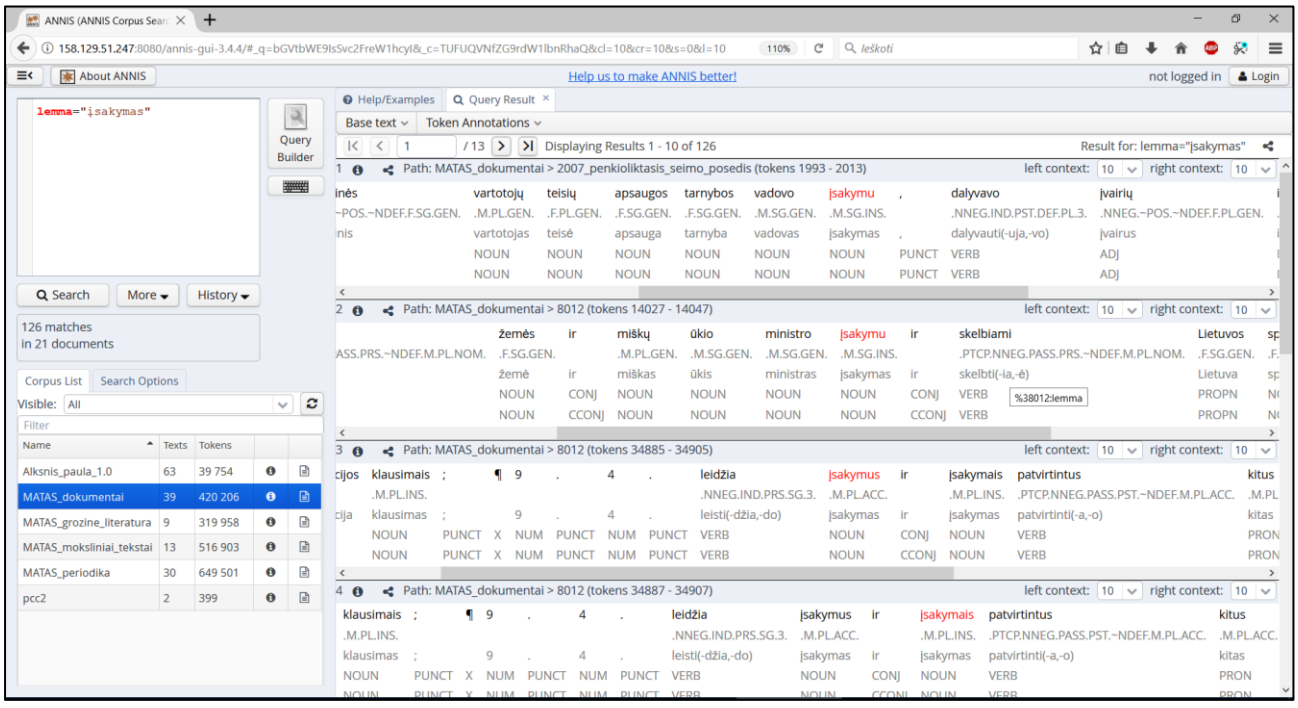
Kitame poskyryje pristatoma duomenų paieška anotuotuose tekstynuose naudojant ANNIS sistemą. Tai vienas iš iššūkių naudojant anotuotus tekstynus, nes duomenų paieška tokiuose tekstynuose, kur yra daugybė pažymų ir net keli anotavimo lygmenys, turi būti įvairialypė. Dėl šios priežasties buvo pasirinktas Humboldtų universiteto (Vokietija) tyrėjų sukurtas įrankis ANNIS.

## 3. Paieška per ANNIS sistemą

Naudoti ANNIS įrankį nėra labai sudėtinga (plačiau žr. Zeldes 2016). Reikia pasirinkti anotuotą tekstyną (MATO tekstyno dalį (šis tekstynas suskirstytas į dokumentus, grožinę literatūrą, mokslinius tekstus ir periodiką) arba ALKSNIO tekstyną (jis įvardytas taip: Alksnis_paula_1.0) ir pateikti užklausą. Užklausų struktūra aprašyta 3.1 ir 3.2 poskyriuose.

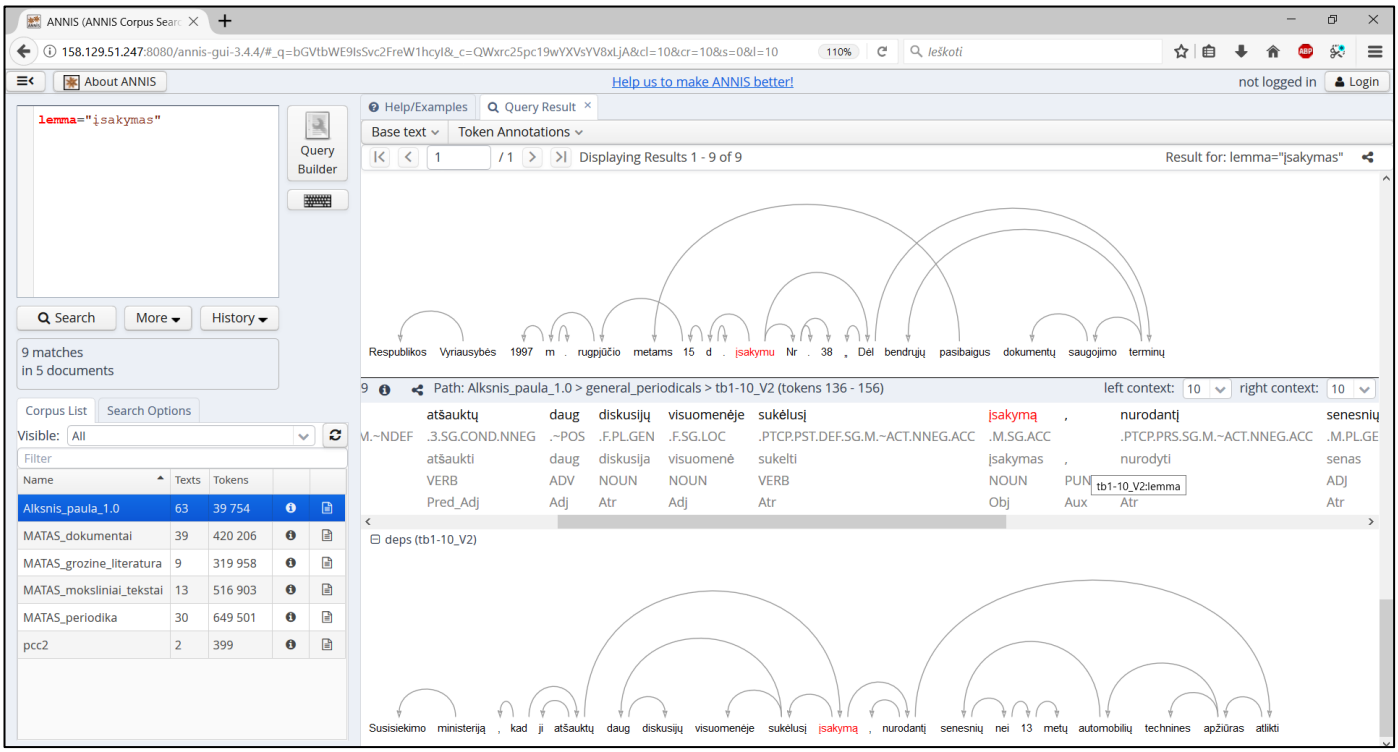
ANNIS pateikia atitinkamas konkordanso eilutes su susijusia informacija (žr. 16 pav.). Konkordanso eilutę galima išplėsti nuo 5 iki 25 žodžių (įskaičiuojami ir skyrybos ženklai) tiek iš kairės, tiek iš dešinės; taip pat galima eksportuoti rezultatus.

Abiejuose tekstynuose naudojamos gramatinės pažymos, sudarytos pagal Leipcigo glosavimo pažymas[21]. Yra pridėtos kelios pažymos, kurių nėra minėtose pažymose, pvz., ~COMP reiškia aukštesnįjį laipsnį. Prieš tokias pažymas rašomas tildės ženklas. Kalbos dalys nurodomos pagal *Universal Dependency* formatą.




16 PAV. Paieškos per ANNIS įrankį rezultatų fragmentas
MATO dokumentų dalyje ieškant lemos *įsakymas*

[21] Prieiga internete: https://www.eva.mpg.de/lingua/resources/glossing-rules.php. Kituose anotuotuose tekstynuose dažnai naudojamos *Universal Dependency* (UD) pažymos (žr. http://universaldependencies.org/; taip pat žr. 18 pav.). Kaip rašyta pirmame skyriuje, UD pažymos labai ilgos (pvz.: galininko linksnis, moteriškoji giminė ir daugiskaita nurodoma taip: case=accusative|gender=feminine|number=plural), jos užima per daug vietos ANNIS sąsajoje ir dėl to nepatogu skaityti informaciją.




17 PAV. Paieškos per ANNIS įrankį rezultatų fragmentas

ALKSNIO tekstyne ieškant lemos *įsakymas*

(kad sintaksiniai ryšiai būtų pažymėti rodyklėmis, reikia paspausti pliusą prie *deps*)

## 3.1. Paprastoji paieška

Vieno požymio paieška vadinama paprastąja. Pateikiamas anotavimas (gramatinių kategorijų pavadinimai, naudojamos gramatinės pažymos ir pan.) priklauso nuo kiekvieno tekstyno struktūros, o ne nuo ANNIS įrankio.

Atkreipiame dėmesį, kad formuojant užklausą reikia skirti didžiąsias ir mažąsias raides.

### 3.1.1. Tikslių žodžių formų (simbolių eilučių) paieška

Paprastoji paieška vykdoma nurodant kategorijos pavadinimą, lygybės ženklą ir kategorijos vertę, įrašytą angliškose kabutėse, pvz.:

- lemma="kalba" (tokia užklausa reiškia, kad ieškoma lemos *kalba*);
- pos="ADJ" (ieškoma būdvardžių);
- gram=".M.SG.GEN." (ieškoma tam tikrų gramatinių kategorijų, šiuo atveju vienaskaitos vyriškąja gimine kilmininko forma pavartotų žodžių).

Ieškant konkrečios žodžio formos tekste, reikia nurodyti kategorijos pavadinimą *tok* arba tiesiog ieškomą formą įrašyti kabutėse, pvz.:

- tok="pasakė" arba "pasakė".

Abiejuose anotuotuose tekstynuose galima ieškoti lemų, kalbos dalių, gramatinių kategorijų, konkrečių žodžių formų, o ALKSNYJE dar galima ieškoti sintaksinių pažymų, pvz.:

- syfun="Atr" (ieškoma atributų).

ALKSNYJE taip pat dar galima ieškoti priklausomybių santykių (*deps*), bet jų ieškoma pagal kitą principą (žr. 3.2.3).

### 3.1.2. Simbolių struktūrų paieška

ANNIS leidžia atlikti paiešką naudojant reguliariąsias išraiškas (angl. *regular expressions)*; jos žymimos specialiais simboliais. Paieška reguliariosiomis išraiškomis iš esmės užrašoma taip pat kaip tikslių žodžių formų paieškos užklausa, tik reikia angliškas kabutes pakeisti pasviraisiais brūkšniais.

3.1.2.1. Bet kokių simbolių paieška

Svarbiausias specialus simbolis yra taškas, jis pakeičia bet kurį vieną simbolį (raidę, skaitmenį ir pan.), pvz.:

- lemma=/pl.t.s/ (ieškoma, pvz., *platus*, *plotis*, *plitus*);
- /p.sak.s/ arba tok=/p.sak.s/ (ieškoma, pvz., *pasakys*, *pasakos*, *pasakas*, *pasakęs, posakis*);
- /201./ arba tok=/201./ (ieškoma, pvz., 2010, 2011, 2012...).

Kadangi taškas gali pakeisti bet kurį ženklą, paties taško simbolio reikia ieškoti naudojant kombinaciją pasvirasis kairinis brūkšnys + taškas (\.), pvz.:

- gram=/\..\.SG\.GEN\./ (ieškoma, pvz., *.M.SG.GEN.*, *.F.SG.GEN.,* ...).

3.1.2.2. Vieno iš kelių simbolių paieška

Vietoj visiškai neapibrėžtų simbolių paieškos užklausoje galima nurodyti keletą konkrečių ieškomų simbolių. Jie rašomi laužtiniuose skliaustuose, pvz.:

- lemma=/pl[ao]t.s/ (ieškoma, pvz., *platus*, *plotis,* bet ne *plitus*);
- /t[ei]lp./ arba tok=/t[ei]lp./ (ieškoma, pvz., *telpa*, *tilpo*, bet ne *talpa*);
- /201[0124]/ arba tok=/201./ (ieškoma *2010, 2011, 2012, 2014*);

- gram=/\.[MFN]\.SG\.GEN\./ (ieškoma *.M.SG.GEN.*, *.F.SG.GEN.*, *.N.SG.GEN.*).

Atkreipiame dėmesį, kad šiuo būdu ieškoma pagal vieną simbolį iš nurodytų simbolių rinkinio: /dirb[aius]/ leidžia ieškoti *dirba*, *dirbi*, *dirbu*, *dirbs,* bet ne *dirbau* arba *dirbsi*.

### 3.1.3. Kartojimo operatoriai

Galima kartoti (įprastus ir specialius) simbolius, pvz.:

- lemma=/..važiuoti/ (ieškoma, pvz., *nuvažiuoti*, *išvažiuoti, apvažiuoti, atvažiuoti, nevažiuoti*);
- /dirb[ao][mt]e/ arba tok=/dirb[ao][mt]e/ (ieškoti *dirbame*, *dirbate, dirbome, dirbote*).

Taip ieškomų simbolių skaičius yra pastovus, pvz., du simboliai pateiktuose pavyzdžiuose (arba [ao][mt]), taigi pagal užklausą lemma=/..važiuoti/ negausite rezultatų su lema *įvažiuoti*. Daugiau galimybių yra su kartojimo operatoriais, jų reikšmės tokios:

- ? – tam tikras simbolis pavartotas vieną kartą arba nė karto;
- + – tam tikras simbolis pavartotas ne mažiau kaip vieną kartą (t. y. vieną, du, tris, keturis ir daugiau kartų);
- * – tam tikras simbolis pavartotas n kartų (t. y. gali būti visai nepavartotas, pavartotas vieną, du ir daugiau kartų)

Šie operatoriai padeda ieškoti simbolių, einančių po jų arba prieš juos, pvz.:

- /šild?o/ (ieškoma *šilo* arba *šildo*);
- /Ma+u/ (ieškoma *Mau*, *Maau*, *Maaau*, *Maaaau* ir t. t.);
- /oi*/ (ieškoma *o*, *oi*, *oii*, *oiii* ir t. t.).

Anksčiau minėti operatoriai yra suderinami su bet kokio ar vieno iš kelių simbolių paieška, pvz.:

- lemma=/.*važiuoti/ (ieškoma, pvz., *važiuoti*, *išvažiuoti, įvažiuoti, nuvažiuoti*, *pravažiuoti, nepravažiuoti*);
- /šauk[aiu]*/ arba tok=/šauk[aiu]*/ (ieškoma, pvz., *šauk, šauki, šaukia*, *šaukiu*);

- /šauk[aiu]+/ arba tok=/šauk[aiu]+/ (ieškoma, pvz., *šauki, šaukia*, *šaukiu,* bet ne *šauk*);
- gram=/.*\.F\..*/ (ieškoma seka *.F.* tarp bet kurių kitų simbolių sekų).

Paminėtina, kad tokios struktūros, kaip [aiu]* arba [aiu]+, neturi įtakos ieškomų simbolių tvarkai. Vadinasi, pateikiami rezultatai, kuriuose yra *a* arba *i,* arba *u* + *a;* arba *i,* arba *u* + *a;* arba *i,* arba *u*... Taigi atpažįstami tokie simbolių deriniai: *aaa*, *i*, *uuuuu*, *uiua*, *uaa* ir t. t.

#### 3.1.4. Alternatyvų paieška

Vertikalusis brūkšnys (|) leidžia ieškoti alternatyvų, pvz.:

- lemma="negeras" | lemma="geras";
- /d(au|ū)ž.*/ arba tok=/d(au|ū)ž.*/ (ieškoma, pvz., *daužo*, *dūžta*, *daužymas*);
- gram=/.*\.~(COMP|SUP)\..*/ (ieškoma .~COMP. ir .~SUP . tarp bet kurių simbolių sekų);
- "tik" | "tiktai" arba tok="tik" | tok="tiktai" (ieškoma *tik* ir *tiktai*).

#### 3.1.5. Neigiamai suformuotos užklausos

Nors tokia galimybė naudingesnė kombinuojant požymius sudėtinėse paieškose, bet ir paprastosiose paieškose galima nurodyti, ko nereikia ieškoti, pvz.:

- pos!="NOUN" (ieškoma visų žodžių, kurie nėra daiktavardžiai);
- tok!=/.*[ąęeėįyouųū]/ (ieškoma visų žodžių formų, kurios nesibaigia balse).

Jei nesirenkama sudėtinės paieškos užklausa (žr. 3.2), neigiamai suformuotos paprastosios užklausos gali sukelti problemų, ypač jei apima daug žodžių (kaip pirmuoju atveju – daiktavardžių), tad šią užklausą reikia naudoti atsargiai.

### 3.2. Sudėtinė paieška

Sudėtinė paieška sudaroma iš kelių paprastųjų paieškų. Dar reikalinga papildoma dalis, kuri aprašo santykius tarp paprastųjų paieškų (nurodytų pagal poziciją). Kiekviena paprastosios paieškos dalis sujungiama ampersendo (&) ženklu.

### 3.2.1. Požymių kombinacijos vienam žodžiui

Pagal paprastąją paiešką ieškoma to paties žodžio, o sąsaja tarp paprastųjų paieškų išreiškiama kombinavimo operatoriumi (_=_), pvz.:

- "mano" & pos="PRON" & #1 _=_ #2 (ieškoti įvardžio *mano*, bet ne veiksmažodžio *manyti* esamojo laiko trečiojo asmens).

Tokia užklausos formuluotė reiškia, kad pirma paprastoji paieška (#1, t. y. "mano" arba tok="mano") ir antra paprastoji paieška (#2, t. y. pos="PRON") susijusi su tuo pačiu žodžiu.

### 3.2.2. Žodžių kombinacijos

Jeigu remiantis paprastąja paieška suformuluojama užklausa skirtingiems žodžiams, kombinavimo operatorius keičiamas sekos operatoriumi (tašku), pvz.:

- lemma="pilnas" & gram=/.*\.GEN\..*/ & #1 .1 #2 (ieškoma kilmininko linksniu pavartoto žodžio po žodžio *pilnas*).

Skaičius 1 po taško reiškia, kad #2 tiesiogiai (be įsiterpimo) eina po #1. Skaičius 2 nurodytų, kad įsiterpia vienas žodis, 3 – du žodžiai ir t. t. Vietoj .1 galima rašyti tašką be jokio skaičiaus (lemma="pilnas" & gram=/.*\.GEN\..*/ & #1 . #2). Galima patikslinti intervalą su kableliu, pvz., .2,4 (su įterpimu tarp vieno ir trijų žodžių). Jeigu įterpimo dydis nesvarbus, reikia parašyti tašką ir žvaigždutę (.*).

Vienoje sudėtinėje užklausoje galima naudoti sekos ir kombinavimo operatorius, pvz.:

- lemma="pilnas" & gram=/.*\.GEN\..*/ & pos!="ADJ" & #1 .1 #2 & #2 _=_ #3 (ieškoma kilmininko formos žodžių, kurie nėra būdvardžiai, po žodžio *pilnas*).

### 3.2.3. Sintaksinių priklausomybių paieška

Priklausomybių struktūrų irgi galima ieškoti per sudėtinę paiešką su tiesioginio valdymo operatoriumi ->, pvz.:

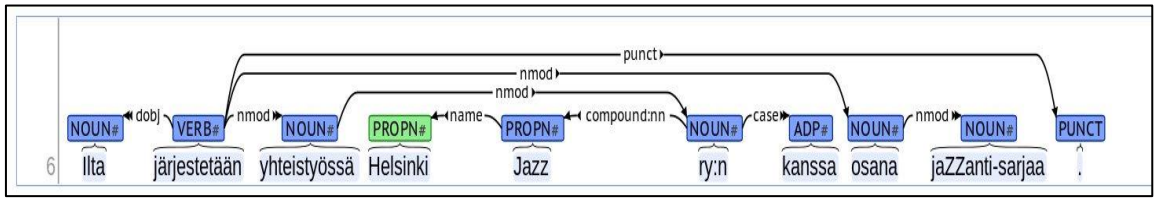
- pos="NOUN" & pos="ADJ" & #1 ->deps #2 (ieškoma daiktavardžių, kurie valdo būdvardžius, pvz., *daugiabutis namas*);
- pos="VERB" & gram!=/.*\.NEG\.*/ & gram=/.*\.GEN\..*/ & syfun="Obj" & #1 _=_ #2 & #3 _=_ #4 & #1 ->deps #3 (ieškoma teigiamųjų veiksmažodžių, kurie valdo kilmininko linksniu išreikštus papildinius, pvz., *laukti klausimų*).

## APIBENDRINAMOSIOS PASTABOS

Tekstynai MATAS ir ALKSNIS yra svarbūs lietuvių kalbos ištekliai. Jie reikalingi tolesniems kalbos kompiuterizavimo etapams. Iš šių tekstynų galima gauti kiekybinių duomenų apie kalbos dalių, gramatinių formų, sintaksinių funkcijų pasiskirstymą, dažniausias lemas, gramatines formas ir pan. (keletą konkrečių kiekybinių tyrimų galima rasti šiuose darbuose: Rimkutė 2006; Utka 2009; Brokaitė 2017). MATAS naudotas mokant semantika.lt analizatorių (anotavimo kokybei įvertinti).

Ateityje būtina toliau plėtoti automatinę sintaksinę analizę. Pirmiausia būtų galima pagerinti sintaksinio anotavimo kokybę: tikslinti sintaksines funkcijas, papildyti naujomis pažymomis (pvz., tikslinti aplinkybes pagal jų semantiką: vietos, būdo, laiko ir t. t.). Sintaksinio anotavimo pradžioje buvo susidurta su nemažai anotavimo neaiškumų, todėl reikėtų diskutuoti dėl kai kurių sintaksinių atvejų, pvz., dėl sudėtinių tarinių anotavimo, elipsės, bejungtukių dėmenų ir aiškinamųjų konstrukcijų skyrimo, kai kurių kalbos dalių statuso ir pan.

Kita darbo kryptis – didinti patį tekstyną. Reikėtų papildyti žanrines tekstyno dalis, pvz., imti daugiau mokslinių ir mokslo populiarinamųjų tekstų. Be to, tekstyną būtina konvertuoti į plačiai naudojamą tarptautinį *Universal Dependency* formatą (žr. 18 pav.; čia nenurodytos gramatinės pažymos, pavyzdyje sintaksiškai anotuotas suomių kalbos pavyzdys).




18 PAV. *Universal Dependency* formatas

Vienas pagrindinių automatinės sintaksinės analizės tikslų – sukurti statistiniu pagrindu veikiantį sintaksinį analizatorių (angl. *statistically-based parser*), kuris būtų paremtas ALKSNIU kaip sintaksinės analizės etalonu ir būtų naudojamas dideliems tekstų kiekiams anotuoti.

Esant galimybių, sintaksinė analizė gali būti plečiama ir kitomis kryptimis, pvz., teminių vaidmenų anotavimas (angl. *thematic role annotation*). Tai yra tam tikras semantinis vaidmuo, susijęs su valentingumu. Šis lygmuo vidinėje sintaksinio anotavimo sistemoje yra priskirtas kiekvienam žodžiui ir šiuo metu viešai nematomas, nes yra labiau eksperimentinis, iki galo neišplėtotas ir nepatikrintas. Taip pat būtų naudinga sužymėti pastoviuosius junginius (angl. *multi-word*

*expressions*, MWE[22]), kurie apima kolokacijas, frazeologizmus. Iki šiol tekstyne jie nebuvo žymimi, išskyrus morfologines samplaikas, kurių dėmenys sujungti į vieną vienetą (pvz., *kol kas, iš tolo*) ir turi vieną morfologinę bei sintaksinę funkciją (dažniausiai žymimi kaip prieveiksmiai ir aplinkybės). Taip pat planuojama pažymėti anaforinius santykius (kai viename sakinyje pavartotas daiktavardis, o kitame sakinyje (ar kitoje sakinio dalyje) jis pakeičiamas įvardžiu, t. y. pažymėti įvardžio ir juo pakeičiamo daiktavardžio sąsajas).

LITERATŪRA

Bielinskienė A., Boizou L., Kovalevskaitė J., Rimkutė E. 2016: Lithuanian dependency treebank ALKSNIS. – *Human language technologies – the Baltic perspective*: proceedings of the 7th international conference, Baltic HLT 2016. Amsterdam: IOS Press, 107–114. Prieiga internete: http://ebooks.iospress.nl/volumearticle/45523 (žiūrėta 2017 10 15).

Boizou L., Zamblera F. 2014. Syntactic engine for the Lithuanian language. – *Human language technologies – the Baltic perspective*: proceedings of the 6th international conference, Baltic HLT 2014. Amsterdam: IOS Press, 69–74. Prieiga internete: http://ebooks.iospress.nl/publication/38006 (žiūrėta 2017 10 15).

Brokaitė K. 2017: *Tarinio raiška gramatinėmis formomis sintaksiškai anotuotame lietuvių kalbos tekstyne ALKSNIS*. Magistro darbas. Vytauto Didžiojo universitetas.

Dadurkevičius 2017: Lietuvių kalbos gramatika skaitmeniniame atvirojo kodo pasaulyje. – 24-osios Jono Jablonskio konferencijos *Skaitmeniniai kalbos ištekliai jų plėtros kryptys ir panaudos galimybės* pranešimų tezės. 2017 m. rugsėjo 29 d., Vilnius.

Hajič J., Panevová J., Buráňová E., Urešová Z., Bémová A. 1999: *ANNOTATIONS AT ANALYTICAL LEVEL. Instructions for annotators* (11.10.1999), UK MFF ÚFAL Praha. Prieiga internete https://ufal.mff.cuni.cz/pedt2.0/publications/a-man-en.pdf (žiūrėta 2017 11 05).

Kapočiūtė-Dzikienė J., Rimkutė E., Boizou L. 2017: A Comparison of Lithuanian Morphological Analyzers. – 20th International Conference *Text, Speech, and Dialogue* (TSD 2017). Springer International Publishing AG, 47–56.

Muischnek K., Müürisep K., Puolakainen T. 2014: Dependency Parsing of Estonian: Statistical and Rule-based Approaches. – *Human Language Technologies – The Baltic Perspective:* proceedings of the 6th international conference, Baltic HLT 2014. Amsterdam:

[22] Pastoviųjų junginių tyrimams skirtas tarptautinis projektas PARSEME, žr. https://typo.uni-konstanz.de/parseme/.

IOS Press, 111–118. Prieiga internete: http://ebooks.iospress.nl/publication/38013 (žiūrėta 2017 10 10).

Pretkalniņa L., Rituma L., Saulīte B. 2016: Universal Dependency Treebank for Latvian: A Pilot. – Proceedings of the 7th International Conference on *Human Language Technologies – the Baltic Perspective:* proceedings of the 7th international conference, Baltic HLT 2016. Amsterdam: IOS Press, 136–143. Prieiga internete: http://ebooks.iospress.nl/volumearticle/45527 (žiūrėta 2017 10 10).

Rimkutė E. 2006: *Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne.* Daktaro disertacija. Vytauto Didžiojo universitetas.

Utka A. 2009: *Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu*. Kaunas: Vytauto Didžiojo universiteto leidykla. Prieiga internete: http://donelaitis.vdu.lt/publikacijos/Dazninis_zodynas.pdf (žiūrėta 2017 12 07).

Utka A., Bielinskienė A., Boizou L., Kovalevskaitė J., Repečka V., Rimkutė E. 2017: Kalbos technologijos – būtina sąlyga kalbai egzistuoti. *BNS Spaudos centras*. 2017 m. spalio 5, 1–4. Prieiga internete: http://sc.bns.lt/view/item/246397 (žiūrėta 2017 10 20).

Zeldes A. 2016: *ANNIS User Guide* (Version 3.4.3). Prieiga internete: http://corpus-tools.org/annis/resources/ANNIS_User_Guide_3.4.3.pdf (žiūrėta 2017 10 31).

Zinkevičius V. 2000: Lemuoklis – morfologinei analizei. – *Darbai ir dienos*, 24, 245–274. Prieiga internete: http://donelaitis.vdu.lt/publikacijos/zinkevicius.pdf (žiūrėta 2017 10 07).

Zipser F., Romary L. 2010: A model oriented approach to the mapping of annotation formats using standards. – *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Malta. Prieiga internete: http://hal.archives-ouvertes.fr/inria-00527799/en/ (žiūrėta 2017 12 07).

Įteikta 2017 11 06

Priimta 2017 12 20

# LITHUANIAN MORPHOLOGICALLY ANNOTATED CORPUS AND TREEBANK

*Summary*

Annotated corpora are fundamental resources, which are very useful to develop language technology. The size, quality, and structure of such annotated corpora has a direct influence on the development of other tools. This article describes two annotated corpora prepared by the Centre of Computational Linguistics at Vytautas Magnus University: MATAS, a morphologically annotated corpus, and ALKSNIS, a treebank. It mainly discusses the structure and the tag set of both corpora, as well as the annotation procedure and tools. Both corpora are available online through ANNIS interface, therefore the syntax of ANNIS simple and complex requests is summarised for the Lithuanian potential users.

KEYWORDS: corpus, automatic morphological analysis, automatic syntactic analysis, treebank, language technologies.

AGNĖ BIELINSKIENĖ, LOÏC BOIZOU, ERIKA RIMKUTĖ
Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras
V. Putvinskio g. 23-216, 44243 Kaunas
*agne.bielinskiene@vdu.lt*
*erika.rimkute@vdu.lt*
*lboizou@gmail.com*